University of California
Santa Barbara

The Impact of Informed Adversarial Behavior in Graphical Coordination Games

A thesis submitted in partial satisfaction
of the requirements for the degree

Master of Science
in
Electrical and Computer Engineering

by

Brian A. Canty

Committee in charge:

Professor Jason Marden, Chair
Professor Mahnoosh Alizadeh
Professor João Hespanha

June 2019

The Thesis of Brian A. Canty is approved.

Professor Mahnoosh Alizadeh

Professor João Hespanha

Professor Jason Marden, Committee Chair

June 2019

The Impact of Informed Adversarial Behavior in Graphical Coordination Games

by

Brian A. Canty

The completion of this undertaking was made possible with the support of the following parties; my sincerest thanks goes out to you. I would like to dedicate this manuscript to my family, my best friend Ryan, my sweet valid boy Zac, my attorney Bernie, the UCSB EWB Alumni Chapter, the Bath Berds, the Anacapa Boys, the MHS Band Alumni family, and all my friends in Harold Frank Hall.

Acknowledgements

I would like to thank from the bottom of my heart my advisors, Professor Jason Marden and Professor Mahnoosh Alizadeh, for allowing me to undertake the adventure that is research while offering their patient, supportive guidance. Additionally, I would like to thank Professor Philip Brown, University of Colorado Colorado Springs, whose collaboration made this thesis possible.

Curriculum Vitæ

Brian A. Canty

Education

2019	M.S. in Electrical Engineering (Expected), University of California, Santa Barbara.
2017	B.S. in Electrical Engineering, University of California, Santa Barbara.

Publications

Canty, B., Brown, P.N., Alizadeh, M., Marden, J.R. (2018). The Impact of Informed Adversarial Behavior in Graphical Coordination Games. CDC, 2018, 8.

Abstract

The Impact of Informed Adversarial Behavior in Graphical Coordination Games

by

Brian A. Canty

How does system-level information impact the ability of an adversary to degrade performance in a networked control system? How does the complexity of an adversary affect its ability to degrade performance? In this thesis, we focus on these questions in the context of graphical coordination games where an adversary can influence a given fraction of the agents in the system. Focusing on the class of ring graphs, we use potential game and resistance tree arguments to explicitly highlight how knowledge of the graph structure and agent identities can be exploited by an adversary to significantly degrade system performance. We demonstrate how the lack of such knowledge drastically reduces the potential harm an adversary can do to the system. Additionally, we show that the ability to employ more complex strategies enables an adversary to do significantly more harm to the system compared to a less capable adversary.

Contents

Curriculum Vitae **vi**

Abstract **vii**

1 Introduction **1**

2 Model and Summary of Results **4**
2.1 The Model 4
2.2 Log-linear Learning 5
2.3 Adversarial Influence and Summary of Results 6

3 Preliminaries **10**
3.1 Potential Games 10
3.2 Resistance Trees 11
3.3 Adversary Models 12

4 Our Contributions **15**
4.1 The Influence of Informed Adversaries 15
4.2 The Influence of Oblivious Adversaries 21

5 Simulations **23**

6 Conclusions and Future Work **29**

A Stationary Informed Adversaries **30**
A.1 Proof of Theorem 4.1.1 30
A.2 Proof of Theorem 4.1.3: 41

B Mobile Informed Adversaries **42**
B.1 Proof of Theorem 4.1.4 42
B.2 Proof of Theorem 4.1.6 63

C Oblivious Adversaries **66**
C.1 Proof of Theorem 4.2.1 66
C.2 Proof of Theorem 4.2.2 66

Bibliography **68**

Chapter 1

Introduction

A multiagent system can be viewed as a collection of decision-making entities that are pre-programmed with a control strategy that specifies decisions for all potential observations. These observations could convey information regarding the local environment, as well as information regarding the behavior of other agents in the system. Regardless of the specific problem domain and informational characteristics, the underlying goal is to derive agent control policies that ensure that the emergent collective behavior is desirable with respect to a system-level performance metric.

There are several results in the literature on networked control systems that provide strong guarantees on the quality of the stochastically stable states under the condition that all agents follow the prescribed control policies, e.g. consensus and flocking [1, 2], sensor allocation [3, 4], coordination of unmanned vehicles [5], and others. Here, the fact that an agent's control policy is influenced by the behavior of other agents may create risks with regards to adversarial interventions. Accordingly, in this thesis we ask whether an adversary can exploit these interconnections to negatively influence the quality of the emergent collective behavior.

The baseline control strategy that this thesis considers originates from the game theoretic literature on distributed control [6–9]. One approach in this literature that has received signifi-

cant research attention is (i) assigning each agent a local objective function that is equal to the agent's marginal contribution to the true system-level objective and (ii) assigning each agent a probabilistic distributed learning rule known as *log-linear learning* (see Section 2.2) [10–12]. The allure of this approach is that it guarantees that the emergent collective behavior will optimize the system-level objective (in an asymptotic sense) for a broad class of multiagent systems [13]. However, the susceptibility of this approach to adversarial interventions is generally unknown.

This thesis characterizes the impact of adversarial interventions on log-linear learning for a class of games known as graphical coordination games [14, 15]. In a graphical coordination game, each agent selects a convention and derives a benefit for this selection that depends on how many of the agent's neighbors (in a graph theoretic sense) have selected the same convention. We focus on the case where there are only two conventions denoted x and y, and the potential benefit derived from x is strictly greater than the potential benefit from y. Note that this does not imply that an agent should always choose x; rather, an agent's best convention choice relies heavily on the convention choices of the agent's neighbors. It is well-known that if all agents follow the log-linear learning rule in a graphical coordination game, the resulting asymptotic behavior will optimize social welfare irrespective of the underlying graph or the convention choices available to the agents [16].

In this thesis we seek to characterize how an adversary can degrade the quality of the emergent collective behavior associated with log-linear learning in graphical coordination games by posing as one of the agents in the system. Note that the sole adversarial power in this realm is influencing the agents' decisions through influencing their objective functions. These questions were initially posed in [17] with several interesting finds ranging from the susceptibility of certain graph structures to an analysis on various models of adversarial behavior [18]. In general, the questions addressed in [18] focused on whether or not an adversary could employ a strategy that would ensure the emergent collective behavior is complete coordination on the

inferior convention y. In the present paper, we move away from this constraint and focus on adversarial strategies that minimize the system-level objective, i.e., minimize social welfare.

This thesis focuses on ring graphs where an adversary can influence a given fraction of the agents in the system. By influence, we mean that the adversary poses as a neighboring agent that is selecting one of the two conventions, thereby attempting to influence the system's agents' behavior. The main contributions include the following:

— In Theorems 4.1.1 and 4.1.4, we demonstrate how an adversary can exploit knowledge of the graph structure and the agent identities to derive a strategy that leads to significantly worse behavior than the adoption of the inferior convention y. Interestingly, the optimal adversarial strategy involves broadcasting both the desirable x and inferior y conventions to various subsets of agents.

— In Theorems 4.2.1 and 4.2.2, we strip away the adversary's knowledge of graph structure and the agents' identities. In this scenario, we demonstrate that the adversary can never benefit from broadcasting the desirable convention x, and the complete adoption of the inferior convention is a best case scenario for the adversary. Our main findings identify the fraction of adversarial influence that is necessary to accomplish this goal under both a deterministic and random adversarial influence strategy.

The value of this characterization is that it begins to shed light on how the information available to the adversary can influence the adversary's ability to harm the system.

Chapter 2

Model and Summary of Results

2.1 The Model

We consider the framework of graphical coordination games where there exists a set of agents $N = \{1, 2, \ldots, n\}$ that interact with other agents in a pairwise fashion over a given graph. Each agent $i \in N$ has an action set $\mathcal{A}_i = \{x, y\}$ and interacts with a set of neighboring agents $\mathcal{N}_i \subseteq N$. Agent $i \in N$ derives a benefit from each neighbor $j \in \mathcal{N}_i$, and the value of the benefit depends on their respective action choices $a_i, a_j \in \{x, y\}$ according to the following symmetric payoff matrix $V : \{x, y\}^2 \to \mathbb{R}^2$

		Agent j: x	Agent j: y
Agent i	x	$1+\alpha, 1+\alpha$	$0, 0$
	y	$0, 0$	$1, 1$

Payoff Matrix

where the *payoff gain* $\alpha > 0$ is a parameter that captures the benefit of coordinating on action x as opposed to action y. In a graphical coordination game, the utility of agent i, denoted by

$U_i : \mathcal{A} \to \mathbb{R}$, is merely the total payoff derived from the pairwise interactions, i.e.,

$$U_i(a_i, a_{-i}) = \sum_{j \in \mathcal{N}_i} V(a_i, a_j), \tag{2.1}$$

where $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ denotes the set of joint actions and $a_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$ denotes the actions of all agents other than agent i. Finally, we measure the quality of a joint action profile $a \in \mathcal{A}$ as

$$W(a) = \sum_{i \in N} U_i(a). \tag{2.2}$$

It is straightforward to see that for any graph G defined by any neighbor sets $\{\mathcal{N}_i\}_{i \in N}$, the all-$x$ action profile, denoted $\vec{x} := (x, \ldots, x)$, is a (pure) Nash equilibrium and maximizes W. Note that $\vec{y} := (y, \ldots, y)$ is also a Nash equilibrium, but that it does not maximize W; other sub-optimal equilibria may exist as well depending on the graph structure.

2.2 Log-linear Learning

It is well-known that there exist distributed learning dynamics that converge to the action profile that maximizes $W(a)$, irrespective of the graph G and the structure of the agents' action sets $\mathcal{A}_i \subseteq \{x, y\}$ [8,19,20]. For such settings, the all x action profile is not necessarily optimal. A learning algorithm produces a sequence of joint action $a(0), a(1), \ldots, a(t), \ldots$, when the action profile at any time $t > 0$ is chosen according to a given rule. The learning algorithm known as *log-linear learning* [11, 13, 21] chooses an action profile at time t in the following way:

- Select any agent $i \in N$ with uniform probability.

- Agent i selects their action at time t probabilistically according to

$$\mathbf{Pr}\left[a_i(t) = x\right] = \frac{e^{\beta U_i(x, a_{-i}(t-1))}}{e^{\beta U_i(x, a_{-i}(t-1))} + e^{\beta U_i(y, a_{-i}(t-1))}},$$
$$\mathbf{Pr}\left[a_i(t) = y\right] = \frac{e^{\beta U_i(y, a_{-i}(t-1))}}{e^{\beta U_i(x, a_{-i}(t-1))} + e^{\beta U_i(y, a_{-i}(t-1))}},$$

  where $\beta > 0$ is a given algorithm parameter.

- All other agents repeat their previous action, i.e., $a_{-i}(t) = a_{-i}(t-1)$.

For any $\beta > 0$, the log-linear learning process is known to induce an ergodic Markov process over state space $\mathcal{A}$, where the unique stationary distribution $\pi_\beta = \{\pi_\beta^a\}_{a \in \mathcal{A}} \in \Delta(\mathcal{A})$ is

$$\pi_\beta^a = \frac{e^{\frac{\beta}{2} W(a)}}{\sum_{\tilde{a} \in \mathcal{A}} e^{\frac{\beta}{2} W(\tilde{a})}}. \tag{2.3}$$

The limiting distribution $\pi := \lim_{\beta \to \infty} \pi_\beta$ exists and is unique [13]. If an action profile a' is in the support of the limiting distribution (i.e., $\pi^{a'} > 0$), it is known as a *stochastically stable equilibrium* of the game for log-linear learning [22]. In the nominal game described in this thesis, the set of stochastically stable equilibria is precisely the action profiles that maximize W, namely the all-x action profile $\vec{x}$.

2.3 Adversarial Influence and Summary of Results

We consider a scenario where there is an adversary which can impersonate up to $k \leq n$ counterfeit agents in the graphical coordination game, thereby modifying the payoffs received by the system's agents in an effort to influence the system's agents to choose specific actions. Specifically, the adversary can select two sets $S_x, S_y \subseteq N$ with $S_x \cap S_y = \emptyset$, where S_x (respectively S_y) represents the set of agents who are connected to a counterfeit agent that is playing a fixed action x (respectively y). In this revised setting, the utility of any agent $i \in N$

is now of the form

$$\tilde{U}_i(a_i, a_{-i}) = \begin{cases} \sum_{j \in \mathcal{N}_i} V(a_i, a_j) + V(a_i, x) & \text{if } i \in S_x, \\ \sum_{j \in \mathcal{N}_i} V(a_i, a_j) + V(a_i, y) & \text{if } i \in S_y, \\ \sum_{j \in \mathcal{N}_i} V(a_i, a_j) & \text{otherwise.} \end{cases} \tag{2.4}$$

That is, an agent $i \in S_x$ (respectively $i \in S_y$) receives its usual payoffs from its neighbors in $\mathcal{N}_i$ as well as an additional payoff of $1 + \alpha$ if $a_i = x$ (respectively, 1 if $a_i = y$).

In order to compare the performance of configurations on graphs of different sizes, we introduce the *efficiency* η of a configuration. Efficiency is the ratio of the quality of some action profile a to the quality of the optimal configuration on the same graph:

$$\eta(a) = \frac{W(a)}{\max_a W(a)}. \tag{2.5}$$

The optimal configuration for a graph of any size is the all-x state $\vec{x}$. Hence, we have that $\eta(a) = \frac{W(a)}{2(1+\alpha)n}$.

The central question that we seek to address in this thesis is the following: given a graph G and constraints on the level of adversarial influence of the form $|S_x| + |S_y| \leq k \leq n$, what are the adversarial influence sets S_x and S_y that minimize the efficiency η of the stochastically stable state associated with the log-linear learning process operating on the influenced utility functions given in (2.4)?

The main results of the paper include the following:

— In Theorems 4.1.1, 4.1.2, 4.1.4, and 4.1.5 we show how a well-informed adversary with various capabilities can minimize the efficiency of the stochastically stable state in a ring graph. A key feature here is that the adversary optimally influences a small number of agents to play the superior x action and others to play y, as this results in some agents failing to coordinate with one another, driving down the efficiency. Given payoff gain α and a fractional adversarial bud-

get $\gamma = k/n$, an informed adversary with basic capabilities can use Theorem 4.1.1 to determine the minimal efficiency action profile capable of being stochastically stable when the adversary best utilizes its resources. Then, the adversary can use Theorem 4.1.2 to construct the necessary adversary set to accomplish stability of the profile. Theorem 4.1.3 subsequently proves a tight bound on the effectiveness of this influence. Similarly, an informed adversary with complex capabilities can use Theorem 4.1.4 to determine the minimal efficiency action profile capable of being stochastically stable when the adversary best utilizes its resources. Then, the adversary can use Theorem 4.1.5 to construct the necessary adversary sets to accomplish stability of the profile. Theorem 4.1.6 subsequently proves a tight bound on the effectiveness of this influence.

— In Theorem 4.2.1, we show that if the adversary *cannot* observe the indices of the agents, this renders the adversary incapable of targeting specific action profiles. Here, the oblivious adversary's optimal action is simply to attempt to influence the agents to choose the all-y state $\vec{y}$. We show a similar limit on efficiency here, though in this case the adversary can be considerably less effective.

— Theorem 4.2.2 considers a case in which the adversary cannot observe the indices of the agents, and attempts to compensate for this by influencing the agents in a random "mobile" way, randomly changing the influence sets at each time step. Here, we show again that the adversary should never influence agents to play x, and that the adversary's maximum influence is similar to that of the oblivious adversary. However, in this case the adversary's influence is totally independent of the number of agents it can influence. We summarize our results quantitatively in Figure 2.1.

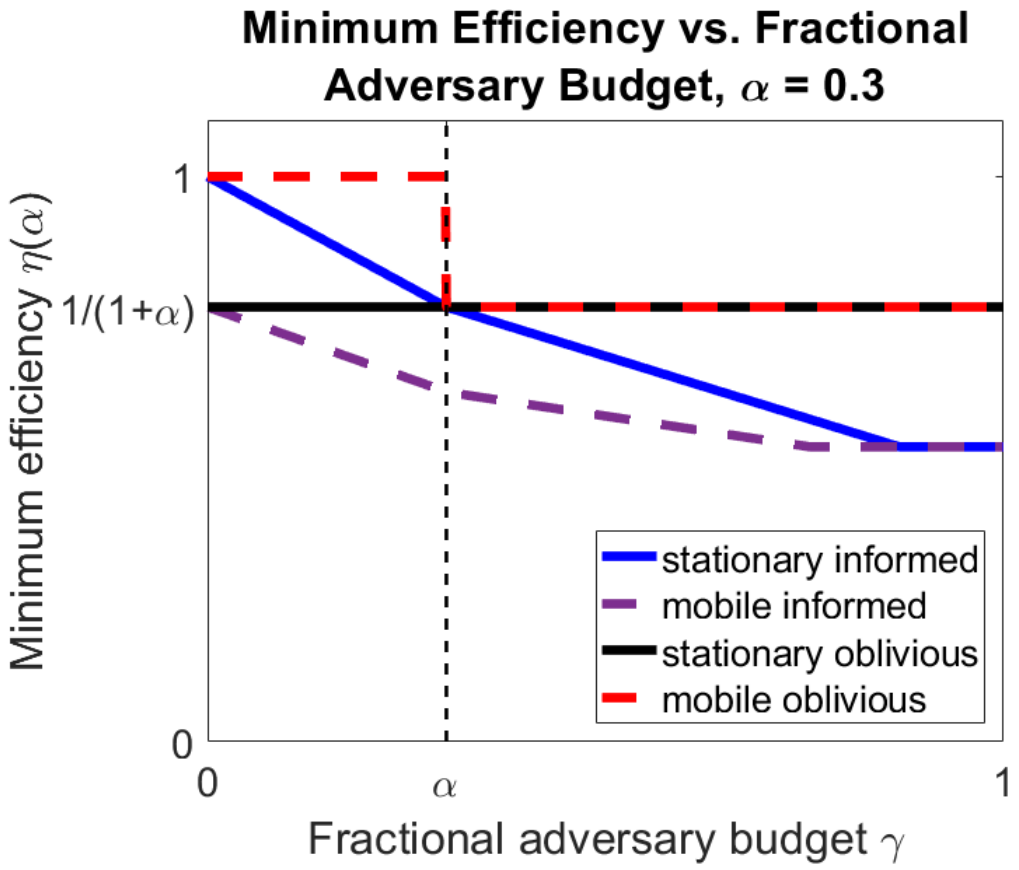

Figure 2.1: The minimum efficiency of any action profile capable of being stabilized in the stationary informed, mobile informed, stationary oblivious, and mobile random oblivious cases for fixed α, where $\gamma = k/n$ is the fractional adversary budget. In the stationary and mobile informed case, the adversary can stabilize a sub-optimal action profile for any $0 < \gamma \leq 1$. In the stationary oblivious case, the adversary is more limited and can stabilize a sub-optimal action profile for any $\alpha < \gamma \leq 1$. The informed adversary can always stabilize some profile with equal or less efficiency than its oblivious counterpart. A random mobile oblivious adversary can stabilize sub-optimal profiles for any value of $\gamma > 0$ and $\alpha < 0.5$.

Chapter 3

Preliminaries

3.1 Potential Games

We begin by reviewing a class of games, termed *potential games* [23], that are intimately related to the class of graphical coordination games.

Definition 3.1.1 *A game $(N, \mathcal{A}, \{U_i\}_{i \in N})$ is a potential game if there exists a potential function $\phi : \mathcal{A} \to \mathbb{R}$ such that for any action profile $a \in \mathcal{A}$, agent $i \in N$, and action $a_i' \in \mathcal{A}_i$,*

$$U_i(a_i, a_{-i}) - U_i(a_i', a_{-i}) = \phi(a_i, a_{-i}) - \phi(a_i', a_{-i}). \tag{3.1}$$

An interesting facet of potential games is that the stationary distribution associated with log-linear learning for any potential game can be expressed in terms of the potential function as

$$\pi^a = \frac{e^{\beta \phi(a)}}{\sum_{\tilde{a} \in \mathcal{A}} e^{\beta \phi(\tilde{a})}} \tag{3.2}$$

for any $\beta > 0$. Hence, the support of the limiting distribution (and thus the set of stochastically stable equilibria) is equal to the set of maximizers of the potential function ϕ, and we will use

this characterization extensively in the forthcoming arguments.

For the graphical coordination game without adversaries specified by agent utilities as defined in (2.1), it is well-known that this game is a potential game with potential function

$$\phi(a) = \frac{1}{2}\sum_{i\in N}\sum_{j\in\mathcal{N}_i} V(a_i, a_j) = \frac{W(a)}{2}, \tag{3.3}$$

irrespective of the underlying graph structure or the parameter α. Accordingly, this fact gives rise to the stationary distribution given in (2.3) by substituting (3.3) into (3.2). On the other hand, the graphical coordination game with adversaries, i.e., agent utilities as defined in (2.4), is also a potential game where the potential function is now of the form

$$\phi(a; S_x, S_y) = \frac{W(a)}{2} + \sum_{i\in S_x} V(a_i, x) + \sum_{i\in S_y} V(a_i, y). \tag{3.4}$$

For this case, the optimizers of (3.4) will now consist of the stochastically stable states.

3.2 Resistance Trees

We now review resistance tree arguments, which must be employed in the case that S_x and S_y are allocated dynamically as functions of the action profile a of the system. In this case, potential game arguments cannot be used to determine the stochastically stable state of the system, and resistance tree arguments must be used instead. For a detailed review of resistance trees, see [22].

Let P^0 be the probability transition matrix for a finite state Markov chain over state space $\mathcal{A}$. Denote the recurrent classes of P^0 as $E_1, E_2, \ldots, E_N$ where each class is a collection of action profiles. For each pair of distinct recurrent classes E_i and E_j, $i \neq j$, an ij-path is defined to be a sequence of distinct states $\zeta = (z_i \to z_2 \to \cdots \to z_n)$ such that $z_1 \in E_i$ and $z_n \in E_j$.

The resistance $r(z_u \to z_v)$ is given by

$$r(z_u \to z_v) = [W(z_v) - W(z_u)]_+ , \tag{3.5}$$

where $[\cdot]_+$ is the projection onto the positive orthant. The resistance of path ζ is given by the sum of the resistance of its edges, such that $r(\zeta) = r(z_1 \to z_2) + r(z_2 \to z_3) + \cdots + r(z_{n-1} \to z_n)$. Let $\rho_{ij} = \min r(\zeta)$ be the least resistance over all ij-paths ζ.

Now construct a complete directed graph with N vertices, one for each recurrent class. The vertex corresponding to class E_j will be called j. The weight on the directed edge $i \to j$ is ρ_{ij}. A tree, T, rooted at vertex j is a set of $N-1$ directed edges such that from every vertex different from j, there is a unique directed path in the tree to j. The resistance of a rooted tree, T, is the sum of the resistances ρ_{ij} on the $N-1$ edges that compose it. The stochastic potential, ψ_j, of the recurrent class E_j is defined to be the minimum resistance over all trees rooted at j. The support of the limiting distribution (and thus the set of stochastically stable equilibria) is given by states contained in the recurrent classes with minimum stochastic potential:

$$\pi_\beta = \{E_i | i \in \arg\min_j \psi_j\}. \tag{3.6}$$

3.3 Adversary Models

To capture the difference in the complexity of adversaries, we will use two different models for determining the abilities of the adversary. A simple adversary may only be able to select fixed adversary sets S_x and S_y and use those sets for all action profiles of the system it encounters. We call such an adversary *stationary,* and use potential game arguments to determine the stochastically stable states of a system influenced by an adversary of this variety. A more complex adversary may be able to take a dynamic approach and adjust the adversary

sets S_x and S_y depending on the action profile it observes. These adversaries employ *policies* $S : \mathcal{A} \to S_x, S_y$ that map each possible action profile to an adversary set, and will be referred to as *mobile* adversaries. To determine the stochastically stable states of a system attacked by a mobile adversary, resistance tree arguments will be used.

Suppose an adversary knows the identities of each agent and can influence each specific agent with a particular action. We call such an adversary *informed.*

Here, for a graph with $|N| = n$, a stationary informed adversary's goal is to select the influence sets $S_x \subseteq N$ and $S_y \subseteq N$, where $S_x \cap S_y = \emptyset$ and $|S_x| + |S_y| \leq k$, that minimize

$$\eta^{\mathrm{SI}}(\alpha, n, k) := \min_{S_x, S_y} \max_{a^* \in \arg\max \phi(a; S_x, S_y)} \eta(a^*). \tag{3.7}$$

Similarly, a mobile informed adversary's goal is to select the adversary policy $S(\alpha)$, where S_x and S_y are functions of the action profile a of the system, that minimize

$$\eta^{\mathrm{MI}}(\alpha, n, k) := \min_{S_x(a), S_y(a)} \max_{a^* \in \{E_i | i \in \arg\min_j \psi_j\}} \eta(a^*). \tag{3.8}$$

We wish to characterize how adversarial knowledge of the graph G impacts the degree to which the adversary can influence the stochastically stable state. To address this, suppose the adversary knows the agents form a ring graph, but cannot observe the identities of the agents. We call such an adversary *oblivious*, and model this by assuming that the adversary cannot directly select influence sets S_x and S_y, but rather can only choose the sizes $k_x := |S_x|$ and $k_y := |S_y|$. A stationary oblivious adversary seeks to minimize

$$\eta^{\mathrm{SO}}(\alpha, n, k) := \min_{\substack{k_x, k_y \in \mathbb{Z}_+ \\ k_x + k_y \leq k}} \max_{\substack{S_x, S_y \subseteq N \\ |S_x| = k_x \\ |S_y| = k_y}} \max_{a^* \in \arg\max \phi(a; S_x, S_y)} \eta(a^*). \tag{3.9}$$

A mobile informed adversary assigns adversary allocations to each set as a function of the

action profile a and seeks to minimize

$$\eta^{\mathrm{MO}}(\alpha, n, k) := \min_{\substack{k_x(a), k_y(a) \in \mathbb{Z}_+ \\ k_x(a)+k_y(a) \leq k}} \max_{\substack{S_x, S_y \subseteq N \\ |S_x|=k_x(a) \\ |S_y|=k_y(a)}} \max_{a^* \in \{E_i | i \in \arg\min_j \psi_j\}} \eta(a^*). \tag{3.10}$$

Our analysis will primarily be concerned with understanding the maximum damage an adversary can cause for large n. As such, we express results in terms of the fraction of agents that the adversary can influence, which we denote by $\gamma := k/n \in [0, 1]$ and call the adversary's *budget*. The minimum efficiency a stationary informed adversary can induce is thus

$$\eta^{\mathrm{SI}}(\alpha, \gamma) := \liminf_{n \to \infty} \ \eta^{\mathrm{SI}}\left(\alpha, n, \lfloor \gamma n \rfloor\right), \tag{3.11}$$

the minimum efficiency a mobile informed adversary can induce is

$$\eta^{\mathrm{MI}}(\alpha, \gamma) := \liminf_{n \to \infty} \ \eta^{\mathrm{MI}}\left(\alpha, n, \lfloor \gamma n \rfloor\right), \tag{3.12}$$

the minimum efficiency a stationary oblivious adversary can induce is

$$\eta^{\mathrm{SO}}(\alpha, \gamma) := \liminf_{n \to \infty} \ \eta^{\mathrm{SO}}\left(\alpha, n, \lfloor \gamma n \rfloor\right), \tag{3.13}$$

and the minimum efficiency a mobile oblivious adversary can induce is

$$\eta^{\mathrm{MO}}(\alpha, \gamma) := \liminf_{n \to \infty} \ \eta^{\mathrm{MO}}\left(\alpha, n, \lfloor \gamma n \rfloor\right). \tag{3.14}$$

Chapter 4

Our Contributions

In this section we characterize the solutions to (3.7), (3.8), and (3.9) in the context of ring graphs with a large number of agents n, where an adversary can influence at most a $\gamma \in [0, 1]$ fraction of agents in the graph. By ring graph, we mean that the neighbor set of each agent $i \in N$ is $\mathcal{N}_i = \{i-1, i+1\}$. All forthcoming arithmetic on agent indices will be $\bmod n$ where appropriate.

4.1 The Influence of Informed Adversaries

Our first result is for the case that the adversary has full knowledge about the graph and the identities of the agents, and thus can solve (3.7).

Theorem 4.1.1 *Given payoff gain α, adversarial budget $\gamma \geq \alpha$, and sufficiently-large n, the*

minimum efficiency an informed adversary can induce (in the sense of (3.11)*) is given by*

$$\eta^{\mathrm{SI}}(\alpha,\gamma) =$$
$$\min_{\ell_{x_1},\ell_{x_2},\ell_{y_1},\ell_{y_2}} \frac{1}{1+\alpha}\left(1+\frac{(2+\alpha)(\frac{s_1}{s_2}-1)+\alpha(\ell_{x_1}-\frac{s_1}{s_2}\ell_{x_2})}{\ell_{x_1}+\ell_{y_1}-\frac{s_1}{s_2}(\ell_{x_2}+\ell_{y_2})}\right)$$
s.t.
$$\ell_{x_1},\ell_{x_2}\geq 2$$
$$\ell_{y_1},\ell_{y_2}\geq \max\left\{\frac{2+\alpha}{1-\alpha},3\right\}$$
$$\ell_{x_\kappa},\ell_{y_\kappa}\in\mathbb{Z}_+$$
$$s_\kappa=\gamma(\ell_{x_\kappa}+\ell_{y_\kappa})-\lfloor\alpha(\ell_{y_\kappa}+1)\rfloor-\left\lfloor\left[\frac{2-\alpha(\ell_{x_\kappa}-1)}{1+\alpha}\right]_+\right\rfloor-4$$
$$s_1\geq 0$$
$$s_2<0 \tag{4.1}$$

where $[\cdot]_+$ *is the projection onto the positive orthant.*

By solving the optimization problem in Theorem 4.1.1, it is also possible to determine the exact optimal structure of the adversary's influence sets S_x and S_y, as described in Theorem 4.1.2:

Theorem 4.1.2 *Given payoff gain* α *and adversarial budget* γ*, let* $\ell^*_{x_1}$*,* $\ell^*_{x_2}$*,* $\ell^*_{y_1}$ *and* $\ell^*_{y_2}$ *be the optimizers of* (4.1)*. For each* $\kappa\in\{1,2\}$*, let* s_κ *be defined in terms of* $\ell^*_{x_\kappa}$ *and* ℓ_{y_κ} *as in* (4.1)*, and choose integers* $r_1,r_2\in\mathbb{Z}_+$ *to satisfy* $r_1s_1+r_2s_2=0$*. Let* $m_{x_\kappa}:=\left\lfloor\left[\frac{2-\alpha(\ell_{x_\kappa}-1)}{1+\alpha}\right]_+\right\rfloor+1$*. The optimal informed adversary policy constructs influence sets* S_x *and* S_y *out of prototypical influence sets (denoted* S^1_x,S^2_x,S^1_y,S^2_y*) that repeat along the ring graph in a specific pattern. For each of the following, let* i *denote an arbitrary agent. First,* $S^\kappa_x=\{i,\ldots,i+m_{x_\kappa}-1\}$ *(if* $m_{x_\kappa}=0$*, then* $S^\kappa_x=\emptyset$*). Similarly, there exists a heuristic for constructing the set* S_y *outlined*

in Lemma A.1.3 in the Appendix. Then repeat each S_x^1 and S_y^1 r_1 times, repeat S_x^2 and S_y^2 r_2 times, and then repeat the overall pattern continually. This results in influencing the agents with the following overall pattern of actions:

$$\underbrace{\underbrace{(\overbrace{xx\cdots x}^{\ell_{x_1}^* \text{ agents}})(\overbrace{yy\cdots y}^{\ell_{y_1}^* \text{ agents}})}_{\text{repeated } r_1 \text{ times}}\underbrace{(\overbrace{xx\cdots x}^{\ell_{x_2}^* \text{ agents}})(\overbrace{yy\cdots y}^{\ell_{y_2}^* \text{ agents}})}_{\text{repeated } r_2 \text{ times}}\cdots}_{\text{repeated indefinitely}} \tag{4.2}$$

We illustrate the results of these theorems in the following example:

Example 4.1.1 *Consider a ring graph with $n = 33$ agents as depicted in Figure 4.1. Let payoff gain $\alpha = 1/3 - \epsilon$, where $\epsilon \to 0^+$. Suppose the adversary can attack with $k = 22$ so that $\gamma = 2/3$. In this case, the solution to the optimization problem* (4.1) *is $\ell_{x_1}^* = 2$, $\ell_{y_1}^* = 11$, $\ell_{y_2}^* = 5$, and $\ell_{x_2}^* = 2$, with the associated $s_1 = 2/3$ and $s_2 = -4/3$. A choice of r_1 and r_2 which satisfies $r_1 s_1 + r_2 s_2 = 0$ is to set $r_1 = 2$ and $r_2 = 1$.*

Thus, to cause the maximum harm to the system, the adversary should influence the agents in two repeating patterns with actions of the form $x \ldots xy \cdots y$; the first should have lengths $\ell_{x_1}^* = 2$, $\ell_{y_1}^* = 11$, with $\lfloor \alpha(\ell_{y_1}^* + 1) \rfloor + \left\lfloor \left[\frac{2-\alpha(\ell_{x_1}^* - 1)}{1+\alpha} \right]_+ \right\rfloor + 4 = 8$ agents influenced and be repeated $r_1 = 2$ times, and the second should have lengths $\ell_{x_2}^* = 2$, $\ell_{y_2} = 5$ with $\lfloor \alpha(\ell_{y_2}^* + 1) \rfloor + \left\lfloor \left[\frac{2-\alpha(\ell_{x_2}^* - 1)}{1+\alpha} \right]_+ \right\rfloor + 4 = 6$ agents influenced and be repeated $r_2 = 1$ times. These patterns, as well as the exact influence sets, are all depicted visually in Figure 4.1.

The efficiency of the resulting stochastically stable action profile is $\eta^{\mathrm{SI}}(1/3, 2/3) = \frac{7}{11} \approx 0.636$, considerably lower than the efficiency of the all-y state on the same graph: $\eta(\vec{y}) = 3/4 = 0.75$.

The minimum efficiency $\eta^{\mathrm{SI}}(\alpha, \gamma)$ in general admits no closed-form expression. However, by relaxing the integrality constraint in (4.1), we can obtain a lower bound on the efficiency resulting from the influence of a stationary informed adversary.

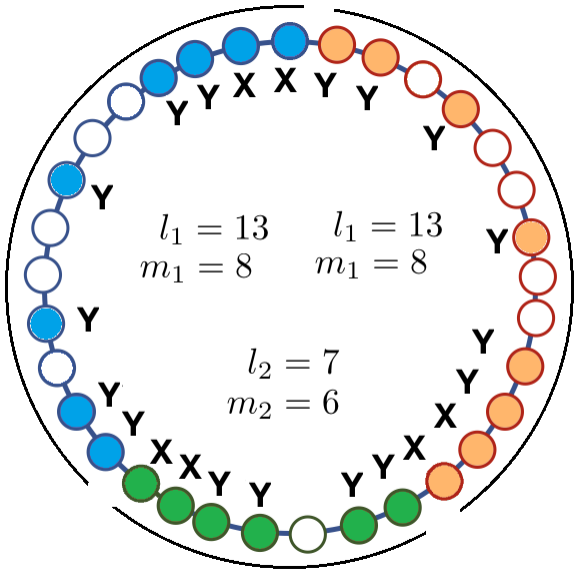

Figure 4.1: Example 4.1.1: An informed adversary's optimal attack on a 33-agent ring when $\alpha = 1/3 - \epsilon$, $\epsilon \to 0^+$ and $k = 22$ (i.e., $\gamma = 2/3$). The optimal attack partitions the agents into two groups of 13 (attacking each of these with 8 adversaries) and one of 7 (attacking this with 6 adversaries). The colors show the three different groups of agents under influence, the bold letters depict the actions being influenced by the adversary, and the open circles indicate agents that are not being attacked.

Theorem 4.1.3 *The lower bound on the achievable efficiency η^* of a system with some payoff gain α and adversarial budget γ is given by:*

$$\eta^{\mathrm{SI}^*}(\alpha,\gamma) = \begin{cases} 1 - \frac{1}{1+\alpha}\gamma, & 0 \le \gamma \le \alpha \\ \frac{2+2\alpha+\alpha^2-\gamma-\alpha\gamma}{2+3\alpha+\alpha^2}, & \alpha < \gamma < b(\alpha) \\ \frac{-2-2\alpha+\alpha^2}{(-4+\alpha)(1+\alpha)}, & b(\alpha) \le \gamma \le 1 \end{cases} \tag{4.3}$$

where

$$b(\alpha) = \frac{-2(2+\alpha^2)}{(-4+\alpha)(1+\alpha)}.$$

Our next set of results are for the case where the adversary is mobile and has full knowledge about the graph and the identities of the agents. Although the proof techniques are different, the results in this case are very similar to the stationary informed case except the key difference

is that a mobile informed adversary can cause the same amount of damage with a significantly lower adversary budget. We present a slightly modified version of 4.1.1 for mobile informed adversaries:

Theorem 4.1.4 *Given payoff gain α, adversarial budget $\gamma \geq \alpha$, and sufficiently-large n, the minimum efficiency a mobile informed adversary can induce (in the sense of* (3.11)*) is given by the solution to the following optimization problem:*

$$\eta^{\mathrm{MI}}(\alpha,\gamma) =$$

$$\min_{\ell_{x_1},\ell_{x_2},\ell_{y_1},\ell_{y_2}} \frac{1}{1+\alpha}\left(1+\frac{(2+\alpha)(\frac{s_1}{s_2}-1)+\alpha(\ell_{x_1}-\frac{s_1}{s_2}\ell_{x_2})}{\ell_{x_1}+\ell_{y_1}-\frac{s_1}{s_2}(\ell_{x_2}+\ell_{y_2})}\right)$$

s.t.

$$\ell_{x_1},\ell_{x_2} \geq 2$$

$$\ell_{y_1},\ell_{y_2} \geq \max\left\{\frac{2+\alpha}{1-\alpha},3\right\}$$

$$\ell_{x_\kappa},\ell_{y_\kappa} \in \mathbb{Z}_+$$

$$s_\kappa = \begin{cases} \gamma(\ell_{x_\kappa}+\ell_{y_\kappa})-4 & \textit{if } \ell_{x_\kappa}=2 \textit{ and } \alpha<0.5 \\ \gamma(\ell_{x_\kappa}+\ell_{y_\kappa})-2 & \textit{else} \end{cases}$$

$$s_1 \geq 0$$

$$s_2 < 0. \tag{4.4}$$

By solving the optimization problem in Theorem 4.1.4, it is also possible to determine the exact optimal structure of the adversary's policy $S : \mathcal{A} \to S_x, S_y$, as described in Theorem 4.1.5:

Theorem 4.1.5 *Given payoff gain α and adversarial budget γ, let $\ell^*_{x_1}$, $\ell^*_{x_2}$, $\ell^*_{y_1}$ and $\ell^*_{y_2}$ be the optimizers of* (4.4)*. For each $\kappa \in \{1,2\}$, let s_κ be defined in terms of $\ell^*_{x_\kappa}$ and ℓ_{y_κ} as in* (4.4)*,*

and choose integers $r_1, r_2 \in \mathbb{Z}_+$ to satisfy $r_1 s_1 + r_2 s_2 = 0$. Let a_k be the action profile obtained by appending ℓ_{x_k} agents who play x to ℓ_{y_k} agents who play y. Then, an action profile a can be constructed by appending together r_1 instances of a_1 and r_2 instances of a_2. The adversary should then use an aggressive policy, outlined in Definition B.1.5 in Appendix B, with target a in order to stabilize the action profile with minimal possible efficiency.

The minimum efficiency $\eta^{\mathrm{MI}}(\alpha, \gamma)$ in general admits no closed-form expression. However, by relaxing the integrality constraint in (4.1), we can obtain a lower bound on the efficiency resulting from the influence of a mobile informed adversary.

Theorem 4.1.6 *The lower bound on the achievable efficiency η^{MI^*} of a system with some pay-off gain α and adversarial budget γ is given by:*

$$\eta^{\mathrm{MI}^*}(\alpha,\gamma) = \begin{cases} \begin{cases} \frac{2-\gamma(1+\alpha)}{2(1+\alpha)} & 0 \le \gamma \le b(\alpha) \\ m(y-b(\alpha)) + \eta_1^{\mathrm{MI}^*}(\alpha, b(\alpha)) & b(\alpha) < \gamma < c(\alpha) \\ \frac{4-2c(\alpha)+\alpha c(\alpha)}{4(1+\alpha)} & c(\alpha) \le \gamma \le 1 \end{cases} & \alpha \le \frac{1}{2}, \\ \begin{cases} \frac{2-2\gamma+\alpha\gamma}{2(1+\alpha)} & 0 < \gamma \le b(\alpha) \\ \frac{2-2b(\alpha)+\alpha b(\alpha)}{2(1+\alpha)} & 0 < \gamma \le b(\alpha) \end{cases} & \textit{else}, \end{cases} \tag{4.5}$$

where

$$m = \frac{\eta_2^{\mathrm{MI}^*} - \eta_1^{\mathrm{MI}*}}{c(\alpha) - b(\alpha)},$$

$$\eta_1^{\mathrm{MI}^*}(\alpha, b(\alpha)) = \frac{2 - b(\alpha)(1+\alpha)}{2(1+\alpha)},$$

$$\eta_2^{\mathrm{MI}^*}(\alpha, c(\alpha)) = \frac{4 - 2c(\alpha) + \alpha c(\alpha)}{4(1+\alpha)},$$

$$b(\alpha) = \begin{cases} \frac{2\alpha}{1+3\alpha} & \alpha \leq \frac{1}{4}, \\ \frac{2\alpha(1-\alpha)}{1+\alpha+\alpha^2} & \frac{1}{4} < \alpha \leq \frac{1}{2}, \\ \frac{2(1-\alpha)}{4-\alpha} & \textit{else}, \end{cases}$$

and

$$c(\alpha) = \begin{cases} \frac{4}{5} & \alpha \leq \frac{1}{4}, \\ \frac{4(1-\alpha)}{4-\alpha} & \textit{else}. \end{cases}$$

4.2 The Influence of Oblivious Adversaries

When the adversary has no knowledge of the indices of the agents in the graph, the adversary effectively cannot employ specifically-targeted influence sets such as those described in Theorem 4.1.2. Instead, an oblivious adversary solves the optimization problem (3.9) by specifying merely the *number* of agents it influences with each action. That is, an oblivious adversary chooses $k_x, k_y \in \mathbb{Z}_+$ to minimize the efficiency of stochastically-stable equilibria in worst case over feasible influence sets satisfying $|S_x| = k_x$ and $|S_y| = k_y$.

Theorem 4.2.1 *Given payoff gain α and adversary budget γ, the minimum efficiency a stationary oblivious adversary can induce (in the sense of* (3.13)*) is*

$$\eta^{\mathrm{SO}}(\alpha, \gamma) = \begin{cases} \frac{1}{1+\alpha} & \alpha < \gamma, \\ 1 & \alpha \geq \gamma. \end{cases} \tag{4.6}$$

This is achieved by choosing $k_x = 0$ and $k_y = \lfloor \gamma n \rfloor$.

Theorem 4.2.1 shows that an adversary who cannot observe agent indices should *never* influence any agent to play the x action. Thus, we see that an informed adversary can leverage its

detailed information to incentivize detailed, heterogeneous, highly-inefficient configurations; an oblivious adversary must be content with incentivizing the all-y configuration $\vec{y}$.

Similarly, in the case of mobile oblivious adversaries, the adversary should never influence an agent to play x. This allows us to leverage known results for mobile adversaries to state the following theorem:

Theorem 4.2.2 *For any payoff gain α and adversary capability $k \in \{1, \ldots, n-1\}$, when agent indices are unknown to the adversary, the optimal adversary allocation (in the sense of* (3.8)*) has $k_x = 0$ and $k_y = k$. The resulting efficiency is*

$$\eta^{\mathrm{MO}}(\alpha, n, k) = \begin{cases} \frac{1}{1+\alpha} & \alpha < \frac{1}{2} \\ 1 & \alpha \geq \frac{1}{2}. \end{cases} \tag{4.7}$$

Note that (4.7) gives the influenced efficiency as independent of k (that is, independent of γ), which suggests an important qualitative difference between deterministic (stationary) adversaries and probabilistic (mobile) ones: an adversary with low γ can influence behavior far more effectively if it is capable of employing a mobile, randomized strategy when compared with a fixed, deterministic one.

Chapter 5

Simulations

The results obtained from running the optimizations posed in Theorems 4.1.1 and 4.1.4 with varying α and γ values have been compiled in Table 5.1 and 5.2 respectively. Our results have also been compiled graphically in Figures 5.1, 5.2, and 5.3 where α is held at a constant value while γ varies.

There are several notable trends to discuss. When $0 < \alpha < 1$, achievable system efficiency decreases as adversarial budget γ increases. Additionally, the value of α only has a small effect on the minimum achievable efficiency relative to the role that γ has. Regardless of the value of α, if $\gamma = 1$, the system operates at nearly half efficiency with lower efficiency achieved at higher α values.

The efficiency of the most damaging action profile a mobile informed adversary can stabilize will always be the same or less than the most damaging profile a stationary informed adversary can stabilize. Thus, mobile adversaries are more capable of decreasing system efficiency in all scenarios. However, when $\gamma \to 1$, both types of informed adversary achieve the same minimum efficiency. A mobile informed adversary cannot stabilize any profile with efficiency worse than the worst efficiency action profile a stationary informed adversary can stabilize with $\gamma = 1$, however the mobile informed adversary is capable of stabilizing this

profile at a significantly lower γ value.

When α is small, there is a noticeable gap between the minimum realizable efficiency and the minimum efficiency bound. This discrepancy stems from the integer constraint set forth in Theorems 4.1.1 and 4.1.4.

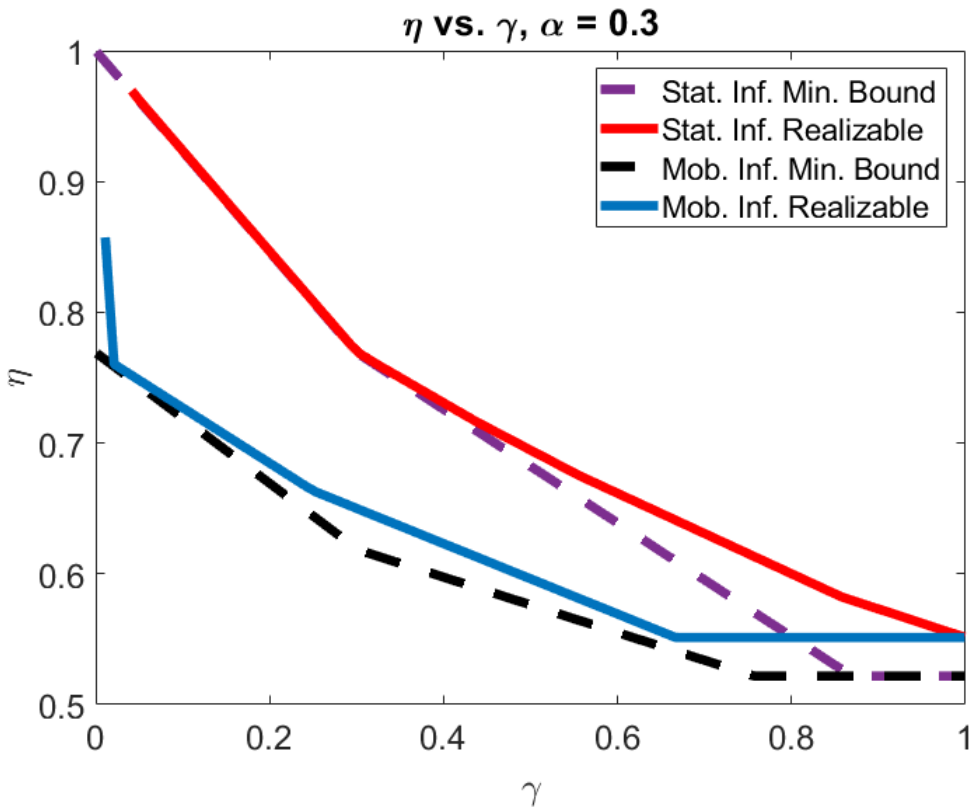

Figure 5.1: Minimum achievable efficiency as a function of fractional adversarial budget γ, $\alpha = 0.3$.

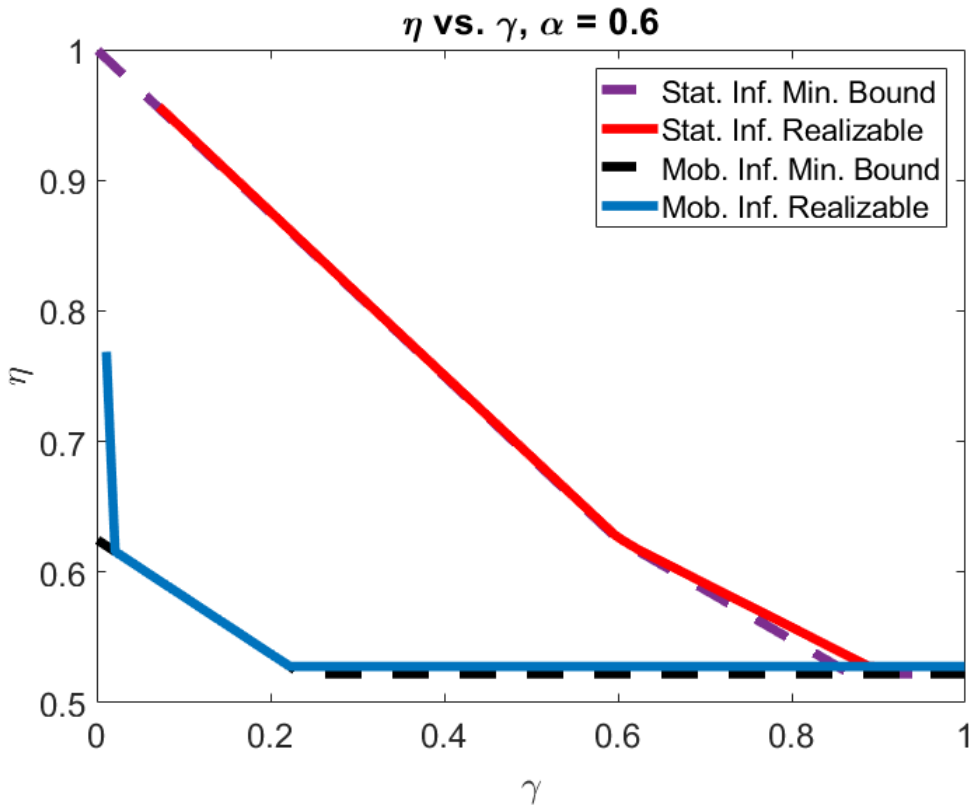

Figure 5.2: Minimum achievable efficiency as a function of fractional adversarial budget γ, $\alpha = 0.6$.

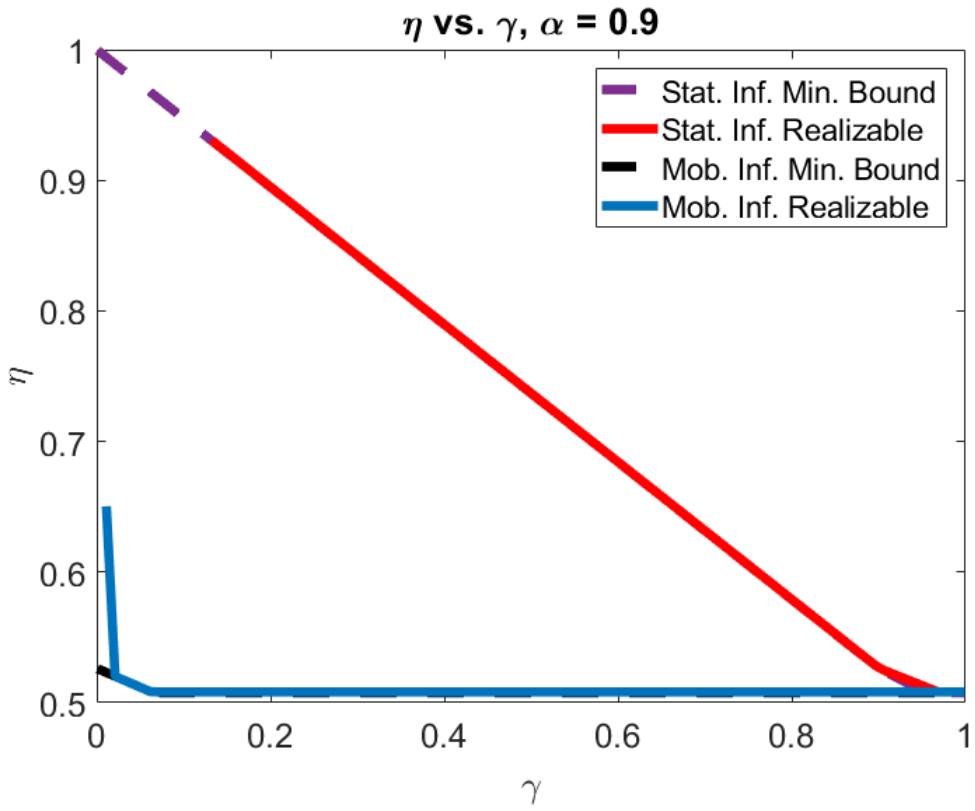

Figure 5.3: Minimum achievable efficiency as a function of fractional adversarial budget γ, $\alpha = 0.9$.

Stationary Informed Adversary							
γ	ℓ_{x_1}	ℓ_{y_1}	ℓ_{x_2}	ℓ_{y_2}	r	η	η^*
$\alpha = 0.3$							
0.2	12	102	95	102	1.43	0.845	0.844
0.5	4	5	4	12	0.42	0.693	0.680
0.8	2	5	4	5	0.15	0.598	0.549
1	2	4	-	-	-	0.552	0.522
$\alpha = 0.6$							
0.2	102	52	102	32	0.54	0.874	0.873
0.5	5	107	94	107	0.30	0.685	0.684
0.8	2	7	2	37	0.13	0.555	0.545
1	2	7	-	-	-	0.528	0.522
$\alpha = 0.9$							
0.2	202	60	202	30	0.23	0.894	0.893
0.5	7	230	200	230	12.6	0.734	0.734
0.8	5	230	57	230	0.93	0.575	0.575
1	2	30	-	-	-	0.508	0.508

Table 5.1: Minimum efficiency configuration characterizations for a stationary informed adversary at various α and γ values. The value r is the number of instances of the second segment that will appear for every instance of the first segment. η is the achieved efficiency, while η^* is the lower bound on efficiency.

Mobile Informed Adversary							
γ	ℓ_{x_1}	ℓ_{y_1}	ℓ_{x_2}	ℓ_{y_2}	r	η	η^*
$\alpha = 0.3$							
0.2	4	4	4	9	0.914	0.688	0.673
0.5	2	4	4	4	0.526	0.598	0.577
0.8	2	4	-	-	-	0.551	0.522
1	2	4	-	-	-	0.551	0.522
$\alpha = 0.6$							
0.2	2	7	2	18	0.54	0.541	0.541
0.5	2	7	-	-	-	0.528	0.522
0.8	2	7	-	-	-	0.528	0.522
1	2	7	-	-	-	0.528	0.522
$\alpha = 0.9$							
0.2	2	30	-	-	-	0.508	0.508
0.5	2	30	-	-	-	0.508	0.508
0.8	2	30	-	-	-	0.508	0.508
1	2	30	-	-	-	0.508	0.508

Table 5.2: Minimum efficiency configuration characterizations for a mobile informed adversary at various α and γ values for a mobile informed adversary. The value r is the number of instances of the second segment that will appear for every instance of the first segment. η is the achieved efficiency, while η^* is the lower bound on efficiency.

Chapter 6

Conclusions and Future Work

This thesis has investigated the susceptibility of distributed learning rules to adversarial manipulation and shown analytic bounds on how much more damage a well-informed adversary can cause than an oblivious one. Additionally, we have shown that an informed adversary that is able to deploy complex strategies can do significantly more damage in many scenarios compared to an adversary who can only deploy simple, fixed strategies. Finally, we have shown that an adversary can employ a randomized approach to compensate for low capabilities.

Future work will involve posing the problem explicitly as a game between the adversary and the system operator and understanding how the operator's knowledge of the adversary's capability can inform various approaches to security.

Appendix A

Stationary Informed Adversaries

A.1 Proof of Theorem 4.1.1

Before delving into the proof, we begin by introducing some convenient notation. First, for any set $S \subseteq N$ and action profile $a \in \mathcal{A}$, let $a_S = \{a_i : i \in S\}$ correspond to the action choices associated with the group S in the action profile a. Accordingly, we extend the definition of $\phi(a)$ (and similarly $\phi(a; S_x, S_y)$) to be restricted to the set S as

$$\phi(a_S) = \frac{1}{2} \sum_{i \in S} \sum_{j \in \mathcal{N}_i \cap S} V(a_i, a_j). \tag{A.1}$$

Lastly, note that any action profile consists of alternating contiguous x and y segments of varying lengths. Without loss of generality, we consider the case where $a_1 = x$ and $a_n = y$, unless of course the action profile is of the form $a = (x, \ldots, x) = \vec{x}$ or $a = (y, \ldots, y) = \vec{y}$. Accordingly, we will represent an action profile by the length of the contiguous components, i.e., $(\ell_x^1, \ell_y^1, \ell_x^2, \ell_y^2, \ldots, \ell_x^m, \ell_y^m)$, where each $\ell_x^k, \ell_y^k \geq 1$ and there are m different segments in the corresponding ring graph. In relation to the action profile a, this means that $a_i = x$ for all $i \in \{1, \ldots, \ell_x^1\}$, $a_i = y$ for all $i \in \{\ell_x^1 + 1, \ldots, \ell_x^1 + \ell_y^1\}$, and so forth. By definition, we have

that $\ell_x^1+\cdots+\ell_x^m+\ell_y^1+\cdots+\ell_y^m = n$. Finally, we let $\vec{\ell}_x = \{\ell_x^1, \ldots, \ell_x^m\}$ and $\vec{\ell}_y = \{\ell_y^1, \ldots, \ell_y^m\}$.

We begin with a useful lemma that provides a characterization of the potential difference between two different action profiles.

Lemma A.1.1 *Let a and a' be any two action profiles. If $a_i = a'_i$ for all players in a given a set $S \subseteq N$, then the potential difference satisfies*

$$\phi(a) - \phi(a') = \phi(a_S) - \phi(a'_S) + \sum_{i\in S}\sum_{j\in\mathcal{N}_i/S}\left(V(a_i,a_j) - V(a_i,a'_j)\right), \tag{A.2}$$

where a_{-S} is shorthand notation for $a_{N\setminus S}$.

Proof: First, note that the potential function for any a and set $S \subseteq N$ can be expressed as

$$\phi(a) = \phi(a_S) + \sum_{i\in S}\sum_{j\in\mathcal{N}_i/S} V(a_i,a_j) + \phi(a_{-S}), \tag{A.3}$$

which follows from the fact that

$$\sum_{i\in S}\sum_{j\in\mathcal{N}_i\setminus S} V(a_i,a_j) = \sum_{i\in N/S}\sum_{j\in\mathcal{N}_i\cap S} V(a_i,a_j) \tag{A.4}$$

from the symmetry in V and the fact that $j \in \mathcal{N}_i$ implies $i \in \mathcal{N}_j$. The result follows by noting that $\phi(a_S) = \phi(a_S)'$ for the considered action profiles. ■

Our next lemma rules out certain configurations as candidates for stochastically stable equilibria.

Lemma A.1.2 *Let a be any action profile with corresponding length vectors $\vec{\ell}_x$ and $\vec{\ell}_y$ and adversarial sets S_x and S_y. If a is stochastically stable, then $\ell_x^k \geq 2$ and $\ell_y^k \geq 3$ for all $k \in \{1,\ldots,m\}$.*

Proof: We will prove this claim by demonstrating that any instance xyx, $xyyx$, or yxy can never be stabilized for any choice of S_x and S_y. We begin by focusing on the case xyx,

meaning that there exists a set of agents $S = \{i, i+1, i+2\}$ such that $a_S = (x, y, x)$. Now, consider the profile $a' = (a'_S, a_{-S})$ where $a'_S = (x, x, x)$. From (A.2), we know that

$$\phi(a, S_{xy}) - \phi(a', S_{xy}) = \phi(a_S, S_{xy}) - \phi(a'_S, S_{xy}), \tag{A.5}$$

where we use the shorthand notation $S_{xy} = (S_x, S_y)$. First, we know that for any sets S_x and S_y,

$$\phi(a, S_{xy}) \leq 1 + 2(1 + \alpha), \tag{A.6}$$

which is achieved when $\{i, i+2\} \in S_x$ and $\{i+1\} \in S_y$. Likewise, we know that for any sets S_x and S_y,

$$\phi(a', S_{xy}) \geq 4(1 + \alpha)). \tag{A.7}$$

Consequently, $\phi(a, S_{xy}) < \phi(a', S_{xy})$. Hence a is not stochastically stable. Similar arguments can be constructed for the other cases as well. ■

Our next lemma begins the process of understanding how many adversaries are necessary to stabilize y-segments of varying length. Before stating the lemma, we introduce the following notation. For a given action profile a with accompanying length vectors $\vec{\ell}_x$ and $\vec{\ell}_y$, let $Q(\ell_x^k) \subseteq N$ capture the ℓ_x^k contiguous player indices in the k-th x component. $Q(\ell_y^k) \subseteq N$ is defined identically.

Lemma A.1.3 *Let a be any action profile with corresponding length vectors $\vec{\ell}_x$ and $\vec{\ell}_y$ and adversarial sets S_x and S_y. If a is the unique stochastically stable, then for any $k \in \{1, \ldots, m\}$*

$$|Q(\ell_y^k) \cap S_x| + |Q(\ell_y^k) \cap S_y| \geq \lfloor \alpha(\ell_y^k + 1) \rfloor + 3. \tag{A.8}$$

Furthermore, the S_x and S_y that minimizes $|Q(\ell_y^k) \cap S_x| + |Q(\ell_y^k) \cap S_y|$ and makes a the unique stochastically stable state achieves this bound with equality.

Proof: Let $Q(\ell_y^k) = \{i, i+1, \ldots, i+\ell_y^k - 1\}$ for simplicity. Note that $a_{i-1} = a_{i+\ell_y^k} = x$ by definition. Accordingly, for ease of presentation we will drop player indices and express such a string as merely $xy^{\ell_y^k}x$. Note that the potential function associated with this segment is

$$\phi\left(xy^{(\ell_y^k)}x; S_{xy}\right) = (\ell_y^k - 1) + |Q(\ell_y^k) \cap S_y|. \tag{A.9}$$

Alternatively, the potential of this section if all members of group $Q(\ell_y^k)$ switched their choice from y to x would be

$$\phi\left(xx^{(\ell_y^k)}x; S_{xy}\right) \geq (1+\alpha)(\ell_y^k + 1), \tag{A.10}$$

where we get equality if $Q(\ell_y^k) \cap S_x = \emptyset$. If a is the unique stochastically stable state, then we know that

$$\phi\left(xy^{(\ell_y^k)}x; S_{xy}\right) > \phi\left(xx^{(\ell_y^k)}x; S_{xy}\right), \tag{A.11}$$

which results in (A.8). We will now switch to establishing sufficiency. That is, we can ensure that the segment $xy^{(\ell_y^k)}x$ is the unique stochastically stable state through the appropriate design of S_x and S_y with the property that $|Q(\ell_y^k) \cap S_x| = 0$ and $|Q(\ell_y^k) \cap S_y| = \lfloor \alpha(\ell_y^k + 1) \rfloor + 3$. Before specifying a particular S_x and S_y, we begin by identifying a series of necessary conditions that ensure the action profile a is the unique stochastically state. We then proceed to demonstrate the sufficiency of these conditions. To that end, the different configurations that we need to consider for the segment $Q(\ell_y^k) = \{i, i+1, \ldots, i+\ell_y^k - 1\}$ are as follows:

- Case #1: $y^{(c)}x^{(d)}y^{(e)}$, $c, d, e > 0$, $c + d + e = \ell_y^k$, which shall be referred to as a cluster of x *within* $Q(\ell_y^k)$,
- Case #2: $x^{(c)}y^{(\ell_y^k - c)}$ or $y^{(c)}x^{(\ell_y^k - c)}$, $0 < c < \ell_y^k - 1$, which shall be referred to as a cluster of x *on the edge of* $Q(\ell_y^k)$,
- Case #3: Some combination of clusters of x within and on the edge of $Q(\ell_y^k)$.

We begin with Case #1. To ensure that the configuration $y^{(\ell_y^k)}$ has strictly higher potential than any cluster of x within $Q(\ell_y^k)$ we will conduct the same potential function analysis done previously. Let $T_1 \subset Q(\ell_y^k)$ be the set of agents that switch to x in the configuration $y^{(c)}x^{(d)}y^{(e)}$. Inspecting the potential function evaluated at these action profiles gives us

$$\begin{aligned} \phi(y^{(\ell_y^k)}) &= \ell_y^k - 1 + |S_y \cap Q(\ell_y^k)| \\ \phi(y^c x^d y^e) &\geq \ell_y^k - d - 2 + (1+\alpha)(d-1) \\ &\quad + |S_y \cap Q(\ell_y^k) \setminus T|. \end{aligned}$$

If $\phi(y^{(\ell_y^k)}) > \phi(y^c x^d y^e)$, then combining the above equations gives us that

$$|S_y \cap T_1| > \alpha(|T_1| - 1) - 2. \tag{A.12}$$

This represents our first condition that needs to be satisfied.

With regards to Case #2, the potential function for the segment $x^{(c)}y^{(\ell_y^k - c)}$ satisfies

$$\phi(x^{(c)}y^{(\ell_y^k - c)}) \geq \ell_y^k - c - 1 + (1+\alpha)c + |S_y \cap Q(\ell_y^k) \setminus T_2|,$$

where $T_2 \subset Q(\ell_y^k)$ is once again the set of agents that switch to x in the configuration $x^{(c)}y^{(\ell_y^k - c)}$. Note that the other symmetric case is identical. If $\phi(y^{(\ell_y^k)}) > \phi(x^{(c)}y^{(\ell_y^k - c)})$, then this gives us

$$|S_y \cap T_2| > \alpha|T_2|. \tag{A.13}$$

This represents our second condition that needs to be satisfied.

With regards to Case #3, if there are some combination of clusters of x within and/or on the edge of $Q(\ell_y^k)$, this will merely require applying the conditions given in (A.12) and (A.13) iteratively over each of the smaller segments that are of Case #1 or Case #2.

We will now provide a minimal construction of the adversary sets S_y and S_x that will ensure that the potential of the segment $xy^{(\ell_y^k)}x$ that satisfies the conditions given in Cases #1-3. First, let $S_x \cap Q(\ell_y^k) = \emptyset$. Define the sets W_1 and W_2 as follows:

$$W_1 = \{w : \lfloor \alpha(w - i + 1) \rfloor - \lfloor \alpha(w - i) \rfloor > 0\}, \tag{A.14}$$

$$W_2 = \{i, w, \ell_y^k - i + 1\}, \tag{A.15}$$

where $w = \max(Q(\ell_y^k)/(W_1 \cap \{\ell_y^k - i + 1\}))$, i.e. the largest index that is neither in W_1 or $\ell_y^k - i + 1$. Then, let $S_y = W_1 \cap W_2$.

This resulting adversary set satisfies the two conditions that have been set forth. In order to show this, we need to determine the minimum number of adversaries that influence any group of agents T_1 within and T_2 on the edge of $Q(\ell_y^k)$.

For groups T_2 that contain the agent i, the number of adversaries that influence the group is given by $\lfloor \alpha |T_2| \rfloor + 1$, which clearly satisfies (A.13).

The number of adversaries k that influence any group T_1 can be lower bounded by taking the difference in the number of adversaries that influence the group T_2' and T_2'', where T_2' is the group that contains agent i and has the same largest index as T_1 and T_2'' is the largest group that contains agent i but does not contain any agents that are in T_1:

$$k \geq \lfloor \alpha(|T_2'| - |T_2''|) \rfloor . \tag{A.16}$$

The size of T_1 is given by taking the difference between T_2' and T_2''. Thus,

$$k \geq \lfloor \alpha(|T_1|) \rfloor . \tag{A.17}$$

This satisfies (A.12) for all T_1.

We can apply a similar procedure to determine how many adversaries influence a group T_2

that contains the agent $\ell_y^k - i + 1$. Groups of T_2 that contain the agent $\ell_y^k - i + 1$ but not agent w will have an adversary influencing each agent. Groups of T_2 that contain the agent $\ell_y^k - i + 1$ and the agent w must be treated as a group T_1 with an additional agent who is influenced by an adversary. In addition, there is another guaranteed adversary on w that is not counted in (A.17). Thus, for any group T_2 that contains agent $\ell_y^k - i + 1$,

$$k \geq \begin{cases} |T_2| & w \notin T_2, \\ \lfloor \alpha(|T_2| - 1) \rfloor + 2 & else. \end{cases} \tag{A.18}$$

Both of these cases satisfy (A.13).

Thus, S_y satisfies all conditions set forth in (A.12) and (A.13). The size of S_y is bounded by $|S_y| \leq \left\lfloor \alpha \ell_y^k \right\rfloor + 3$. This does not necessarily satisfy (A.8), so additional adversaries need to be added to arbitrary indexes within $Q(\ell_y^k)$ until S_y is of sufficient size. ■

Lemma A.1.4 *Let a be any action profile with corresponding length vectors $\vec{\ell}_x$ and $\vec{\ell}_y$ and adversarial sets S_x and S_y. If a is the unique stochastically stable, then for any $k \in \{1, \ldots, m\}$*

$$|Q(\ell_x^k) \cap S_x| + |Q(\ell_x^k) \cap S_y| \geq \left\lfloor \frac{2 - \alpha(\ell_x^k - 1)}{1 + \alpha} \right\rfloor + 1. \tag{A.19}$$

Furthermore, the S_x and S_y that minimizes $|Q(\ell_x^k) \cap S_x| + |Q(\ell_x^k) \cap S_y|$ and makes a the unique stochastically stable state achieves this bound with equality.

Proof: The proof follows similarly to the proof of Lemma A.1.3. Let $Q(\ell_x^k) = \{i, i + 1, \ldots, i + \ell_x^k - 1\}$ for simplicity. Note that $a_{i-1} = a_{i+\ell_x^k} = y$ by definition.

The potential function associated with this segment is

$$\phi\left(yx^{(\ell_x^k)}y; S_{xy}\right) = (1 + \alpha)(\ell_x^k - 1) + |Q(\ell_x^k) \cap S_x|. \tag{A.20}$$

Alternatively, the potential of this section if all members of group $Q(\ell_x^k)$ switched their choice from x to y would be

$$\phi\left(yy^{(\ell_x^k)}y; S_{xy}\right) \geq \ell_x^k + 1, \tag{A.21}$$

where we get equality if $Q(\ell_x^k) \cap S_y = \emptyset$. If a is the unique stochastically stable state, then we know that

$$\phi\left(yx^{(\ell_x^k)}y; S_{xy}\right) > \phi\left(yy^{(\ell_x^k)}y; S_{xy}\right), \tag{A.22}$$

which results in (A.19).

We will now switch to establishing sufficiency. That is, we can ensure that the segment $yx^{(\ell_x^k)}y$ is the unique stochastically stable state through the appropriate design of S_x and S_y with the property that $|Q(\ell_x^k) \cap S_y| = 0$ and $|Q(\ell_x^k) \cap S_x| = \left\lfloor \frac{2-\alpha(\ell_x^k-1)}{1+\alpha} \right\rfloor + 1$. Before specifying a particular S_x and S_y, we begin by identifying a series of necessary conditions that ensure the action profile a is the unique stochastically state. We then proceed to demonstrate the sufficiency of these conditions. To that end, the different configurations that we need to consider for the segment $Q(\ell_x^k) = \{i, i+1, \ldots, i+\ell_y^k - 1\}$ are as follows:

- Case #1: $x^{(c)}y^{(d)}x^{(e)}$, $c, d, e > 0$, $c + d + e = \ell_x^k$, which shall be referred to as a cluster of y *within* $Q(\ell_x^k)$,
- Case #2: $y^{(c)}x^{(\ell_x^k - c)}$ or $x^{(c)}y^{(\ell_x^k - c)}$, $0 < c < \ell_x^k - 1$, which shall be referred to as a cluster of y *on the edge of* $Q(\ell_x^k)$,
- Case #3: Some combination of clusters of y within and on the edge of $Q(\ell_x^k)$.

Configurations that belong to Case #1 and #2 will never be stochastically stable. $Q(\ell_x^k) \cap S_y = \emptyset$, meaning that clusters of agents who play y within and on the side of $Q(\ell_x^k)$ will not satisfy (A.12) and (A.13) respectively.

With regards to Case #3, if there are some combination of clusters of y within and/or on

the edge of $Q(\ell_x^k)$, this will merely require applying the conditions given in (A.12) and (A.13) iteratively over each of the smaller segments that are of Case #1 or Case #2.

Thus, (A.19) is the only condition that must be satisfied and can be done so by adding adversaries to arbitrary indexes within $Q(\ell_x^k)$ until S_x is of sufficient size. ■

Let us define some new notation. Consider a heterogeneous action profile a with partition $\mathcal{Q} = \{Q_1, \ldots, Q_M\}$ and $a_m = x^{(l_{x_m})}y^{(l_{y_m})}$. The vector $\vec{\ell_x} = [\ell_{x_m}]$ is defined such that if there exists a Q_m consisting of l_{x_m} agents that play x attached to l_{y_m} agents that play y in a, then $\ell_{x_m} \in \vec{\ell_x}$ and $\ell_{y_m} \in \vec{\ell_y}$. Moreover, the number of partitions of form $a^m = x^{(\ell_{x_m})}y^{(\ell_{y_m})}$ in a is denoted by r_m, which form the repetition vector $\vec{r}$.

The efficiency of an action profile described by $\vec{\ell_x}$, $\vec{\ell_y}$, and $\vec{r}$ is given by:

$$\eta(\vec{\ell_x}, \vec{\ell_y}, \vec{r}) = \frac{\vec{r}^T\left((1+\alpha)\vec{x} + \vec{y}\right) - (2+\alpha)||\vec{r}||_1}{(1+\alpha)\vec{r}^T(\vec{y}+\vec{x})}. \tag{A.23}$$

For the given $\vec{\ell_x}$ and $\vec{\ell_y}$, define the vector $\vec{s}^T = \{s_1, s_2, \ldots, s_m\}$, where

$$\begin{aligned} s_m = \gamma(\ell_{x_m} + \ell_{y_m}) - \lfloor \alpha(\ell_{y_m} + 1) \rfloor \\ - \left\lfloor \left[\frac{2 - \alpha(\ell_{x_m} - 1)}{1+\alpha} \right]_+ \right\rfloor - 4. \end{aligned} \tag{A.24}$$

For a given payoff gain α and budget γ, s_m is the difference between the adversaries available to a partition described by ℓ_{x_m} and ℓ_{y_m} and the minimum number of adversaries needed to ensure the stochastic stability of that partition.

Lemma A.1.5 *Let action profile a be described by some $\vec{\ell_x}$, $\vec{\ell_y}$, $\vec{r}$, and $\vec{s}$ with $\vec{r}^T\vec{s} > 0$ and efficiency $\eta(\vec{\ell_x}, \vec{\ell_y}, \vec{r})$. If there exists some combination of ℓ_{x_m} and ℓ_{y_m} with $s_m < 0$ and $\eta(x^{(\ell_{x_m})}y^{(\ell_{y_m})}) < \eta(\vec{\ell_x}, \vec{\ell_y}, \vec{r})$, then there exists an action profile with lower efficiency than a with $\vec{r}^T\vec{s} = 0$.*

Proof: Let action profile a be described by some $\vec{\ell_x}$, $\vec{\ell_y}$, $\vec{r}$, and $\vec{s}$ with $\vec{r}^T\vec{s} = b$, $b > 0$ and

efficiency $\eta(a)$. Let ℓ_{x_m} and ℓ_{y_m} describe a segment with $s_m = c$, $c < 0$ and $\eta(x^{(\ell_{x_m})}y^{(\ell_{y_m})}) < \eta(a)$. A new action profile $\tilde{a}$ can be constructed by repeating this additional segment $-\frac{b}{c}$ times each time a is repeated, resulting in the following vectors:

$$\vec{\ell}\,'_x = \begin{bmatrix} \vec{\ell}_x \\ \ell_{x_m} \end{bmatrix}, \quad \vec{\ell}\,'_y = \begin{bmatrix} \vec{\ell}_y \\ \ell_{y_m} \end{bmatrix}, \quad \vec{r}\,' = \begin{bmatrix} \vec{r} \\ -\frac{b}{c} \end{bmatrix}. \tag{A.25}$$

From this, a new $\vec{s}\,'$ can be constructed, and $\vec{r}\,'^T\vec{s}\,' = 0$. The efficiency of $\tilde{a}$ can be expressed as a mediant sum. Let $g(\vec{\ell}_x, \vec{\ell}_y, \vec{r}) = \vec{r}^T\left((1+\alpha)\vec{x} + \vec{y}\right) - (2+\alpha)||\vec{r}||_1$ and $h(\vec{\ell}_x, \vec{\ell}_y, \vec{r}) = \vec{r}^T(\vec{y}+\vec{x})$. Then:

$$\eta(\tilde{a}) = \frac{g(\vec{\ell}_x, \vec{\ell}_y, \vec{r}) - \frac{b}{c}g(\ell_{x_m}, \ell_{y_m}, 1)}{h(\vec{\ell}_x, \vec{\ell}_y, \vec{r}) - \frac{b}{c}h(\ell_{x_m}, \ell_{y_m}, 1)}. \tag{A.26}$$

This is a mediant sum of $\eta(a)$ and $\eta(x^{(\ell_{x_m})}y^{(\ell_{y_m})})$. Since $\eta(x^{(\ell_{x_m})}y^{(\ell_{y_m})}) < \eta(a)$, the resulting mediant sum will also be less than $\eta(a)$. Thus, $\eta(\tilde{a}) < \eta(a)$. ■

Lemma A.1.6 *For a given payoff gain α, the profile of the form $a = x^{(\ell_x)}y^{(\ell_y)}$ that has minimal efficiency and is able to appear in a stochastically stable state is described by:*

$$\ell_x = 2 \quad \ell_y = \min\{l : l \in \mathbb{Z}^+, l \geq \lfloor \alpha(l+1) \rfloor + 3\}. \tag{A.27}$$

Proof: The minimum efficiency segment occurs when the strings of agents playing x and y are as short as possible. The shortest strings can only be stabilized when the agents are under full adversarial influence. The shortest string of x is two agents long when $0 < \alpha < 1$, while the shortest string of y when each agent is influenced is given by

$$\ell_y = \min\{l : l \in \mathbb{Z}^+, l \geq \lfloor \alpha(l+1) \rfloor + 3\}. \tag{A.28}$$

■

Corollary A.1.1 *When adversary capability $\gamma = 1$, The minimum efficiency configuration is given by repeating the minimum efficiency segment outlined in Lemma A.1 indefinitely.*

For any $1 > \gamma > 0$, the minimum efficiency segment will have $s_m < 0$. Hence, to minimize $\eta(a)$ when $1 > \gamma > 0$, we can limit our search to configurations of the form $a = (\vec{\ell_x}, \vec{\ell_y}, \vec{r})$ that minimize $\eta(\vec{\ell_x}, \vec{\ell_y}, \vec{r})$ while utilizing all adversarial influence available, i.e., $\vec{r}^T \vec{s} = 0$.

Lemma A.1.7 *Any heterogeneous minimum efficiency action profile satisfies $|\vec{\ell_x}| \leq 2$ and $|\vec{\ell_y}| \leq 2$, i.e. at most two different lengths of groups of agents playing x and two different lengths of groups of agents playing y will be repeated in the minimum efficiency action profile.*

Proof: To prove this statement, it is sufficient to show that if an action profile a^1 with $\vec{\ell_x}^1 = (\ell_{x_1}, \ell_{x_2}, \ell_{x_3})$, $\vec{\ell_y}^1 = (\ell_{y_1}, \ell_{y_2}, \ell_{y_3})$ and $\vec{s}^1 = (s_1, s_2, s_3)$, $s_1 > 0$, $s_2, s_3 < 0$ there exists another action profile a^2 or a^3 with $\vec{\ell_x}^2 = (\ell_{x_1}, \ell_{x_2})$ and $\vec{\ell_y}^2 = (\ell_{y_1}, \ell_{y_2})$ or $\vec{\ell_x}^3 = (\ell_{x_1}, \ell_{x_3})$ and $\vec{\ell_y}^3 = (\ell_{y_1}, \ell_{y_3})$ that has an efficiency at least as low as that of a^1. Define $p_z = \ell_{y_z} - 1 + (1 + \alpha)(\ell_{x_z} - 1)$ and $\ell_z = \ell_{x_z} + \ell_{y_z}$. We know that for a^1 to be the minimum efficiency profile, we must have $(\vec{r}^1)^T \vec{s}^1 = 0$. Under this condition, we can write:

$$\eta(a^1) = \frac{r_2(p_2 - \frac{s_2}{s_1}p_1) + r_3(p_3 - \frac{s_3}{s_1}p_1)}{(1+\alpha)(r_2(\ell_2 - \frac{s_2}{s_1}\ell_1) + r_3(\ell_3 - \frac{s_3}{s_1}\ell_1))}. \tag{A.29}$$

For the other two action profiles a^2 and a^3, we know that r^2 and r^3 satisfy $(\vec{r}^i)^T \vec{s}^i = 0$. From this, we can write

$$\eta(a^2) = \frac{p_2 - \frac{s_2}{s_1}p_1}{(1+\alpha)(\ell_2 - \frac{s_2}{s_1}\ell_1)},\ \eta(a^3) = \frac{p_3 - \frac{s_3}{s_1}p_1}{(1+\alpha)(\ell_3 - \frac{s_3}{s_1}\ell_1)}. \tag{A.30}$$

We can see that $\eta(a^1)$ is a mediant sum of weighted values $\eta(a^2)$ and $\eta(a^3)$. Hence, either $\eta(a^2)$ or $\eta(a^3)$ is less than or equal to $\eta(a^1)$. This result can be expanded to show that for any stabilizable action profile that is made up of multiple subsets with $s_m \geq 0$ and multiple subsets

with $s_m < 0$, there exists an action profile with lower efficiency made up of just two kinds of the subsets in the original action profile. ■

Thus, the search for a minimal efficiency configuration can be restricted to a search over four lengths (two of segments of x, two of y) that satisfy the constraints on adversary set size and complete adversary utilization. ■

A.2 Proof of Theorem 4.1.3:

When the integer constraint in (4.1) is relaxed, the optimization can be simplified down to one variable. In the original optimization, segments of agents playing y could be one of two different lengths, as well as agents playing x. This condition existed to ensure that all available adversaries were utilized. However, when the integer restriction is lifted, it is possible to always utilize all adversaries when all segments of agents playing y are the same length, and all agents playing x are also the same length. Furthermore, the condition that all adversaries are utilized allows for ℓ_x to be solved for in terms of ℓ_y. When substituted into (A.23), the expression simplifies into a function of α and γ.

The final expression is split into three regimes determined by γ. The first boundary, where $\gamma = \alpha$, arises from the use of the projection onto the positive orthant used to determine the number of adversaries necessary to stabilize a segment of x. Specifically, for values of γ less than α, it is necessary to use an x segment that does not require any adversaries to stabilize. The second boundary at $\gamma = b(\alpha)$ arises from the restriction of one adversary to each agent. At $\gamma > b(\alpha)$, the minimum efficiency segment can be stabilized. ■

Appendix B

Mobile Informed Adversaries

B.1 Proof of Theorem 4.1.4

In the forthcoming section, we will describe action profiles relative to a desired "target" action profile that we wish to be the sole stochastically stable state of the system. First, the desired action profile a will be decomposed into segments of neighboring agents who all play the same action profile. Let $a = \{T_1, T_2, \ldots, T_m\}$, where each T_i is of the form $\{x, x, \ldots, x\}$ or $\{y, y, \ldots, y\}$ and the segments alternate between all playing x and all playing y. Let $Q(i)$ refer to the agents whose actions are described by T_i.

Action profiles a' that differ from the target profile will be described relative to the target. Specifically, the segmentation scheme present in a will be applied to a'. For each $Q(i)$, if the agents all play the same action in a', then that segment is referred to as *homogeneous*. Similarly, if the agents in $Q(i)$ do not all play the same action, segment i in a' is referred to as *heterogeneous*.

To describe the adversary policy that could be used to stabilize a desired target profile a, we propose a policy that comprises of strategies applied to individual segments. Specifically, the proposed policy includes "offensive" and "defensive" strategies that exist for each desired

action of the segment, x or y.

Definition B.1.1 *(Offensive x Strategy). Consider the agents u through v. Let $[p, q]$ be the longest chain of agents playing x within the interval $[u, v]$. A policy $S : \mathcal{A} \to \mathcal{S}_k$ uses an offensive x strategy on the interval $[u, v]$ if it satisfies the following conditions:*

1. *If $a(w) \neq x$, $w \in [u, v]$, no agents in $[u, v]$ play x, then $S_x \cap [u, v] \geq 1$. If agent $u - 1$ or $v + 1$ are playing x, then $u \in S_x$ or $v \in S_x$ respectively. If $i - 1$ and $j + 1$ both play x, it is sufficient to have either $i \in S_x$ or $j \in S_x$.*

2. *Else, if $[p, q] \neq [u, v]$, then $\{p - 1, q + 1\} \cap S_x \geq 1$.*

Definition B.1.2 *(Defensive x Strategy). Consider the agents u through v. Let $[p, q]$ be the longest chain of agents playing x within the interval $[u, v]$. A policy $S : \mathcal{A} \to \mathcal{S}_k$ uses a defensive x strategy on the interval $[u, v]$ if $u - v + 1 < \frac{1}{\alpha}$ and $p \neq q$, then $\{p, q\} \in S_x$.*

Definition B.1.3 *(Offensive y Strategy). Consider the agents u through v. Let $[p, q]$ be the longest chain of agents playing y within the interval $[u, v]$. A policy $S : \mathcal{A} \to \mathcal{S}_k$ uses an offensive y strategy on the interval $[u, v]$ if it satisfies the following conditions:*

1. *If no agents in $[u, v]$ play y, there is at least one adversary in S_y such that $S_y \cap [u, v] \geq 1$. If agent $u - 1$ or $v + 1$ are playing y, then $u \in S_y$ or $v \in S_y$ respectively. If $u - 1$ and $v + 1$ both play y, it is sufficient to have either $u \in S_y$ or $v \in S_y$.*

2. *Else, if $[p, q] \neq [u, v]$, then $\{p - 1, q + 1\} \cap S_y \geq 1$.*

Definition B.1.4 *(Defensive y Strategy). Consider the agents u through v. Let $[p, q]$ be the longest chain of agents playing y within the interval $[u, v]$. A policy $S : \mathcal{A} \to \mathcal{S}_k$ uses a defensive y strategy on the interval $[u, v]$ if $\{p, q\} \in S_y$.*

Definition B.1.5 *(Aggressive Policy). Let action profile a be divisible into "homogeneous segments" such that $a = \{T_1, T_2, \ldots, T_m\}$ and each T_i is of the form $\{x, x, \ldots, x\}$ or $\{y, y, \ldots, y\}$ and the segments alternate between all playing x and all playing y. For some other action profile a', let $a'_{Q(i)}$ refer to the actions of the agents that make up segment i in a'. A policy $S : \mathcal{A} \to \mathcal{S}_k$ is an aggressive policy if the following conditions are satisfied when applying the policy to action profile a':*

- *For the lowest index i where segment i is heterogeneous, if T_i is of the form $\{x, x, \ldots, x\}$, then an offensive and defensive x strategy is applied to $Q(i)$. Else, an offensive and defensive y strategy is applied to $Q(i)$.*

- *If there are no heterogeneous segments, then for the lowest index i where $a'_{Q(i)} \neq T_i$ and T_i is of the form $\{x, x, \ldots, x\}$, an offensive and defensive x strategy is applied to $Q(i)$.*

- *If there are no segments where $a'_{Q(i)} \neq T_i$ and T_i is of the form $\{x, x, \ldots, x\}$, then for the smallest index i where $a'_{Q(i)} \neq T_i$ and T_i is of the form $\{y, y, \ldots, y\}$, then an offensive and defensive y strategy is applied to $Q(i)$.*

- *For any segment where $a'_{Q(i)} = T_i$ and T_i is of the form $\{y, y, \ldots, y\}$, a defensive y strategy is applied to $Q(i)$ if T_i is heterogeneous with only one contiguous segment of agents playing y, or the agents neighboring $Q(i)$ play x.*

- *Let $\hat{T}$ be a vector that contains the lengths of all segments where T_i is of the form $\{x, x, \ldots, x\}$, sorted by increasing length. Let:*

$$\lambda = \underset{j}{argmin}\ j \quad s.t. \quad \sum_{i=1}^{j} \left(|\hat{T}_i| - 2 \right) > \frac{1-\alpha}{\alpha}.$$

Then, a defensive x strategy is applied to the first λ segments where T_i is of form $\{x, x, \ldots, x\}$ and the longest chain of agents playing x is two. If there are more than λ of

these segments, then a defensive x strategy is applied to the λ of them with the smallest length.

We reduce the number of states that are candidates for being the sole stochastically stable state of the system, then quantify the resistance of the minimum resistance paths between remaining action profiles. We will primarily consider paths between action profiles that are "similar"- the action profiles are identical except for the actions of a single segment. Our comparisons will be restricted to action profiles that are made up solely of homogeneous segments according to the segmentation scheme set forth in a target profile a. To compare two action profiles, we will use the following notation. The action profiles will be broken into segments consistent with a using $|$, with $|_x$ ($|_y$) denoting that for the segment to the left of the $|$, $T_i = \{x, x, \ldots, x\}$ ($\{y, y, \ldots, y\}$). $\ldots$ are used in places where the two profiles are identical, and X (Y) is used to denote that all agents in the segment play x (y). For instance, consider the profiles $a' = \{\ldots |X|_x \ldots\}$ and $a'' = \{\ldots |Y|_x \ldots\}$. They are identical action profiles, except the agents in a segment play X in a' and Y in a''. Furthermore, the agents in the segment play X in the target profile a.

Lemma B.1.1 *Consider a system attacked by an aggressive policy S with target a. In this system, all recurrent classes solely contain instances of $\{|X|_x Y|_y X|_x\}$, $\{|Y|_y X|_x Y|_y\}$, $\{|X|_x X|_y X|_x\}$, and $\{|Y|_y Y|_x Y|_y\}$.*

Proof: Consider a system that is attacked by an aggressive policy based off of an action profile a. Let a' be any action profile with at least one heterogeneous segment in it, using the segmentation scheme outlined in a. By definition of an aggressive policy, there will be a heterogeneous segment i that is influenced by either an offensive x or y policy. In this case, there is a path of zero resistance from a' to an action profile a'' that is identical to a', except $a''_{Q(i)} = T_i$. Thus, a' cannot be the sole stochastically stable state of the system.

We will now limit our search for recurrent classes to homogeneous action profiles. Any action profile that contains an instance of $\{|X|_y\ Y|_x\}$ will not be stochastically stable because there exists a zero resistance path to a similar action profile where $\{|X|_y\ Y|_x\}$ is replaced with $\{|X|_y\ X|_x\}$. The remaining segment patterns that can be used to construct valid recurrent classes are $\{|X|_xY|_yX|_x\}$, $\{|Y|_yX|_xY|_y\}$, $\{|X|_xX|_yX|_x\}$, and $\{|Y|_yY|_xY|_y\}$. There exist no zero-resistance transitions from any of these segment patterns, so any recurrent class must be composed solely of segments in these patterns. ■

We will construct paths by stringing together a series of unilateral deviations where an agent switches their selected action in an action profile a'. The resistance of a unilateral deviation that occurs at agent i is given by the resistance function $r(f, g, h)$:

$$r(f,g,h) = \big[gV(f,f) + V(f,h) - (2-g)V(\{x,y\} \setminus f, \{x,y\} \setminus f) - V(\{x,y\} \setminus f, h)\big]_+, \quad \text{(B.1)}$$

where $[\,]_+$ is the positive orthant, $f = a'(i)$, $g = |\{j | j \in \mathcal{N}_i, a'(j) = a'(i)\}|$, and

$$h = \begin{cases} x & \text{if } i \in S_x, \\ y & \text{if } i \in S_y, \\ 0 & \text{else.} \end{cases}$$

.

As shorthand, we will refer to different types of unilateral deviations by describing them using the function r. f is the action agent i plays in a', g is the number of neighbors that also play that action, and h is the action an adversary is broadcasting to agent i (0 if there is no adversary). For instance, a deviation of type $r(x, 1, y)$ refers to a deviation between two action profiles that are identical, except agent i switches from x to y. Agent i has one neighbor

playing x and is influenced by an adversary playing y.

Lemma B.1.2 *Consider a system that is influenced by some aggressive policy S that targets action profile a. Then:*

$$R\left(\{\dots|Y|_x\dots\}\to\{\dots|X|_x\dots\}\right)\leq 1-\alpha. \tag{B.2}$$

Proof: Consider a system that is influenced by some aggressive policy S that targets a. Let a' and a'' be two similar profiles where $a'=\{\dots|Y|_x\dots\}$, $a''=\{\dots|Y|_x\dots\}$, and the segment that differs is segment i. Let $Q(i)=[u,v]$. The first path (denoted by $\to_1$) that will be considered is in the case that $|\{j\in\{u,v\}|a_j=x\}|>0$. then an agent neighboring segment i plays x and there is a series of unilateral transitions from a' to a'' with total resistance given by:

$$R\left(\{\dots|Y|_x\dots\}\to_1\{\dots|X|_x\dots\}\right)= \\ (|Q(i)|-1)r(y,1,S_x)+r(y,0,S_x)=0. \tag{B.3}$$

The second case is where $|\{j\in\{u,v\}|a_j=x\}|=0$. In this case, there exists a second path ($\to_2$) with resistance:

$$R\left(\{\dots|Y|_x\dots\}\to_2\{\dots|X|_x\dots\}\right)= \\ r(y,2,0)+(|Q(i)|-2)r(y,1,S_x)+r(y,0,S_x)=1-\alpha. \tag{B.4}$$

These two paths cover all possible action profiles a' and a'', thus

$$R\left(\{\dots|Y|_x\dots\}\to\{\dots|X|_x\dots\}\right)\leq 1-\alpha. \tag{B.5}$$

■

Lemma B.1.3 *Consider a system that is influenced by some aggressive policy S that targets action profile a. Then:*

$$R(\{\dots|X|_x\dots\} \to \{\dots|Y|_x\dots\}) > 1 - \alpha. \tag{B.6}$$

Proof: Consider a system that is influenced by some aggressive policy S that targets a. Let a' and a'' be two similar profiles where $a' = \{\dots|X|_x\dots\}$, $a'' = \{\dots|Y|_x\dots\}$, and the segment that differs is segment i. Let $Q(i) = [u, v]$, and assume that $a_u, a_v = x$.

All transition paths from a'' to a' will have a higher resistance than $1 - \alpha$. Transition paths can be described as one of the three following types:

- have at least one transition of type $r(x, 2, 0) = 2 + 2\alpha$,
- have at least one transition of type $r(x, 1, x) = 1 + 2\alpha$,
- only contain transitions of type $r(x, 1, 0) = \alpha$.

The first two paths will both have minimum resistance greater than $1 - \alpha$. We will now characterize the minimum number of transitions of type $r(x, 1, 0)$ that occur on paths of the third type.

Under an aggressive policy, defensive strategies will be used on up to λ x segments that only have two neighboring agents playing x. Before a defensive x strategy is used on segment i, $|Q(i)| - 2$ agents must switch from x to y (see figure B.1). To construct a path from a' to a'' that consists solely of transitions of type $r(x, 1, 0)$, these transitions must occur at at least λ different segments before occurring at the desired segment i to ensure that a defensive x strategy is not employed on segment i. The minimum total resistance of these transitions is

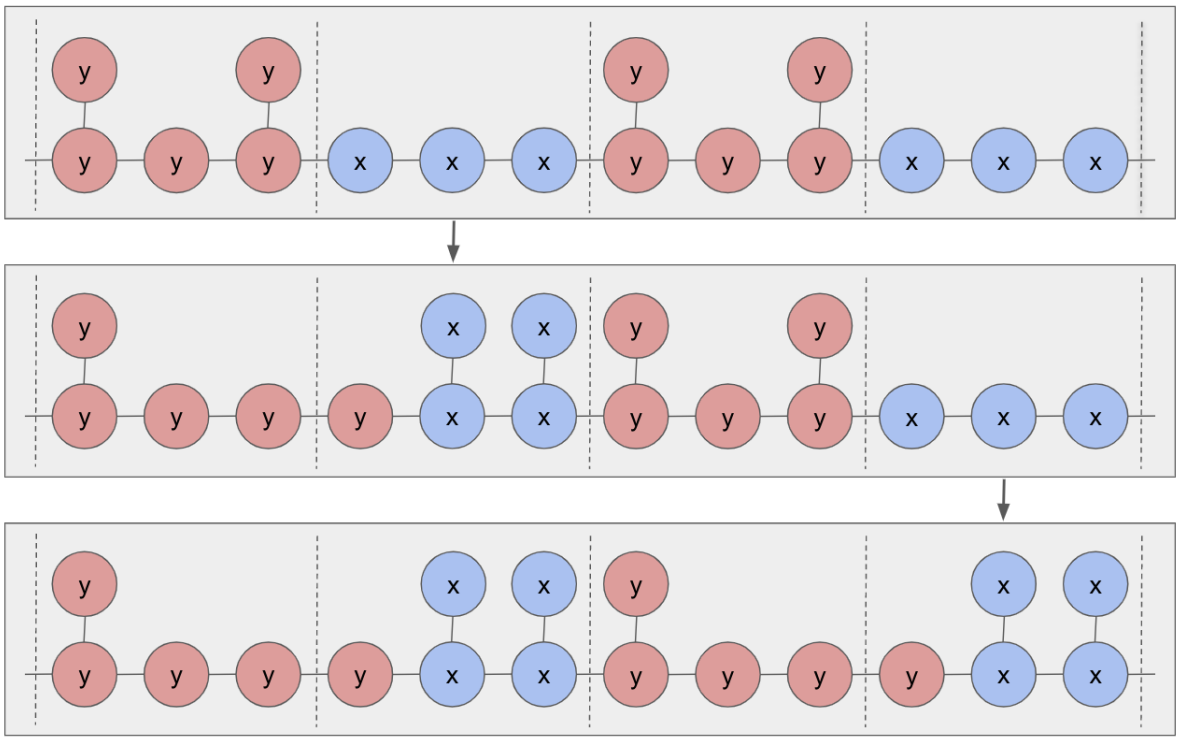

Figure B.1: Adversary assignments under an aggressive policy. No adversaries are deployed to X segments until only two neighboring agents playing x remain.

given by attacking the λ smallest x segments first which yields resistance:

$$R\left(\{\dots|X|_x\dots\} \to_3 \{\dots|Y|_x\dots\}\right) = r(x,1,0)\sum_{i=1}^{\lambda}\left(|\hat{T}_i|-2\right) = \alpha\sum_{i=1}^{\lambda}\left(|\hat{T}_i|-2\right) > \frac{1-\alpha}{\alpha}, \quad \text{(B.7)}$$

which is greater than $1-\alpha$ by the definition of λ. Any paths that target larger x segments will have strictly larger resistance. Thus, all paths will have resistance greater than $1-\alpha$ such that:

$$R\left(\{\dots|X|_x\dots\} \to \{\dots|Y|_x\dots\}\right) > 1-\alpha. \quad \text{(B.8)}$$

■

Lemma B.1.4 *Consider a system that is influenced by some aggressive policy S that targets action profile a. Then:*

$$R\left(\{\ldots|X|_y\ldots\}\rightarrow\{\ldots|Y|_y\ldots\}\right)\leq 1+2\alpha. \tag{B.9}$$

Proof: Consider a system that is influenced by some aggressive policy S that targets a. Let a' and a'' be two similar profiles where $a'=\{\ldots|X|_y\ldots\}$, $a''=\{\ldots|X|_y\ldots\}$, and the segment that differs is segment i. Let $Q(i)=[u,v]$. The first path (denoted by $\rightarrow_1$) that will be considered is in the case that $|\{j\in\{u,v\}|a_j=y\}|>0$. An agent neighboring segment i plays y and there is a series of unilateral transitions from a' to a'' with total resistance given by:

$$R\left(\{\ldots|X|_y\ldots\}\rightarrow_1\{\ldots|Y|_y\ldots\}\right)=(|Q(i)|-1)r(x,1,S_y)+r(x,0,0)=0. \tag{B.10}$$

The second case is where $|\{j\in\{u,v\}|a_j=x\}|=0$. In this case, there exists a second path ($\rightarrow_2$) with resistance:

$$R\left(\{\ldots|X|_y\ldots\}\rightarrow_2\{\ldots|Y|_y\ldots\}\right)= \\ r(x,2,S_y)+(|Q(i)|-2)r(x,1,S_y)+r(x,0,S_y)=1+2\alpha. \tag{B.11}$$

These two paths cover all possible action profiles a' and a'', thus

$$R\left(\{\ldots|Y|_x\ldots\}\rightarrow\{\ldots|X|_x\ldots\}\right)\leq 1+2\alpha. \tag{B.12}$$

■

Lemma B.1.5 *Consider a system that is influenced by some aggressive policy S that targets*

action profile a. Then:

$$R\left(\{\dots|Y|_y\dots\} \to \{\dots|X|_y\dots\}\right) > 1 + 2\alpha. \tag{B.13}$$

Proof: Consider a system that is influenced by some aggressive policy S that targets a. Let a' and a'' be two similar profiles where $a' = \{\dots|X|_y\dots\}$, $a'' = \{\dots|X|_y\dots\}$, and the segment that differs is segment i. Let $Q(i) = [u, v]$, and assume that $a_u, a_v = x$.

All transition paths from a'' to a' will have a higher resistance than $1 + 2\alpha$. Transition paths can be described as one of the three following types:

- no deviations occur involving agents who have two neighbors playing y ($\to_1$),
- only one deviation occurs involving an agent who has two neighbors playing y ($\to_2$),
- more than one deviation occurs involving agents who have two neighbors playing y ($\to_3$).

The resistance of all paths of the first type is given by:

$$R\left(\{\dots|Y|_y\dots\} \to_1 \{\dots|X|_y\dots\}\right) = \\ (|Q(i)| - 1)r(y, 1, S_y) + r(y, 0, S_y) = (|Q(i)| - 1)(1 - \alpha). \tag{B.14}$$

When $|Q(i)| > \frac{2+\alpha}{1-\alpha}$, this resistance will be greater than $1 + 2\alpha$.

We will now limit analysis to lengths $|Q(i)| > \frac{2+\alpha}{1-\alpha}$. In all paths of the second type, one agent with two neighbors playing y will switch to x, leaving two smaller y segments on either end of the agent. Let ℓ_y be the length of the largest segment after this transition, where $\ell_y = \left\lfloor \frac{|Q(i)|}{2} \right\rfloor$. The resistance of all paths of the second type is given by:

$$R\left(\{\dots|Y|_y\dots\} \to_1 \{\dots|X|_y\dots\}\right) = \\ r(y, 2, 0) + (\ell_y)r(y, 1, S_y) + r(y, 0, S_y) = 2 + (1 - \alpha)\ell_y. \tag{B.15}$$

When $|Q(i)| > \frac{2+\alpha}{1-\alpha}$, this resistance will be greater than $1 + 2\alpha$.

The resistance of all paths of the third type will always be greater than 4, since there are at least two deviations involving an agent who has two neighbors playing y. Thus, all possible paths will have a resistance greater than $1 + 2\alpha$. If we change the assumption that the agents surrounding $Q(i)$ play x, the minimum resistance only increases higher than $1+2\alpha$. Therefore,

$$R\left(\{\dots|Y|_y\dots\} \to \{\dots|X|_y\dots\}\right) > 1 + 2\alpha. \tag{B.16}$$

■

To facilitate our classification of recurrent classes, we will introduce some new notation. Every recurrent class can be assigned a level of "disagreement" corresponding to how many segments play different actions relative to their counterparts in a:

$$d(a') = |\{i|a'_{Q(i)} \neq a_{Q(i)}\}|. \tag{B.17}$$

Lemma B.1.6 *In a system that is influenced by some aggressive policy S that targets a, consider the graph G formed by connecting recurrent classes through the minimum resistance edge leaving each class. G consists of disconnected subgraphs $Q_1, Q_2, \dots, Q_m$ such that with each subgraph Q_i:*

1. $\exists\, u, v$ *s.t.* $(u, v), (v, u) \in Q_i$,
2. *if* $(u, v), (v, u) \in Q_i$, *then* $d(u) = \min\{d(j)|j \in Q_i\}$.

Proof: Consider a system that is influenced by some aggressive policy S that targets a. Construct the graph G by connecting recurrent classes through the minimum resistance edge leaving each class.

By Lemma B.1.1, we can limit the action profiles that are recurrent classes to any action profile composed solely of combinations of the patterns $\{|X|_xY|_yX|_x\}$, $\{|Y|_yX|_xY|_y\}$,

$\{|X|_x X|_y X|_x\}$, and $\{|Y|_y Y|_x Y|_y\}$. Recurrent classes can be sorted into two categories: recurrent classes that contain at least one instance of $\{|Y|_y Y|_x Y|_y\}$ and those that do not.

For recurrent classes that contain an instance of $\{|Y|_y Y|_x Y|_y\}$, the least resistance edge leaving the class is to a similar class where $\{|Y|_y Y|_x Y|_y\}$ is replaced with $\{|Y|_y X|_x Y|_y\}$. By Lemma B.1.2, this transition will have a resistance of $1 - \alpha$. Thus, every recurrent class with disagreement z that contains an instance of $\{|Y|_y Y|_x Y|_y\}$ has a minimum resistance edge that leads to a recurrent class of disagreement $z - 1$.

Classes that do not contain an instance of $\{|Y|_y Y|_x Y|_y\}$ can be further broken down into two categories: the action profile $\vec{x}$ and all other remaining recurrent classes. By lemma B.1.4, the minimum resistance path out of $\vec{x}$ has resistance $1 + 2\alpha$ and leads to a class that has disagreement that is one less than $\vec{x}$. All remaining classes will have at least one instance of $\{|X|_x X|_y X|_x Y|_y\}$. There are three possible forms the minimum resistance edge out of these classes will have:

- The minimum resistance edge leads to a similar class where an instance of $\{|X|_x Y|_y X|_x\}$ is replaced with $\{|Y|_y Y|_x Y|_y\}$. By B.1.4 this edge has resistance $1 + 2\alpha$.
- The minimum resistance edge leads to a similar class where an instance of $\{|X|_y X|_x Y|_y\}$ is replaced with $\{|Y|_y Y|_x Y|_y\}$. By B.1.3, this edge has resistance $1 - \alpha < R < 1 + 2\alpha$.
- The minimum resistance edge leads to a similar class where an instance of $\{|Y|_y X|_x Y|_y\}$ is replaced with $\{|Y|_y Y|_x Y|_y\}$. By B.1.3, this edge has resistance $1 - \alpha < R < 1 + 2\alpha$.

If the minimum resistance path is of the first two forms, the edge leads to a class that either has a lower disagreement or has an minimum resistance edge leaving it that connects to a class with lower disagreement. If the minimum resistance path is of the third form, then the edge leads to a class that has higher disagreement. In this case, let the class that has an edge of the third type be u and the class that it is connected to be v.

For each class u that exists, there will be a subgraph Q_i (see figure B.2). All edges within Q_i except the edge leaving u will lead to a class that either has lower disagreement or has an edge connecting that class to a class with lower disagreement, meaning class u has the lowest disagreement of all classes in Q_i. Class v will contain an instance of $\{|Y|_y Y|_x Y|_y\}$, and there will then be a minimum resistance edge connecting v to u. Then,

1. $\exists\, u, v$ s.t. $(u, v), (v, u) \in Q_i$,

2. if $(u, v), (v, u) \in Q_i$, then $d(u) = \min\{d(j) | j \in Q_i\}$.

■

Lemma B.1.7 *The minimum resistance rooted tree in any family Q_i is the minimum resistance subgraph Q_i minus the edge (u, v).*

Proof: Consider a system that is influenced by some aggressive policy S that targets profile a with family Q_i. A rooted tree T can be constructed by taking the minimum resistance subgraph Q_i and removing the edge (u, v). Replacing any of the edges in T other with different edges will result in a rooted tree with potential that is at least as much as T, as all new edges will have at least the same resistance as the original edges in T. Then, the minimum resistance rooted tree in Q_i is the minimum resistance subgraph Q_i minus the edge (u, v). ■

We will refer to the recurrent class a_i^h that the minimum potential rooted tree T_i of family Q_i is rooted in as the "head" class of family Q_i. We will overload the disagreement function such that the disagreement of some rooted tree T_i is given by:

$$d(T_i) = d(a_i^h). \tag{B.18}$$

Lemma B.1.8 *Consider a system that is influenced by some aggressive policy S that targets a. Then, $\nexists\, i$ s.t. $a' \in Q_i : T_i = Y, a^h_{Q(i)} = X, a'_{Q(i)} = Y$.*

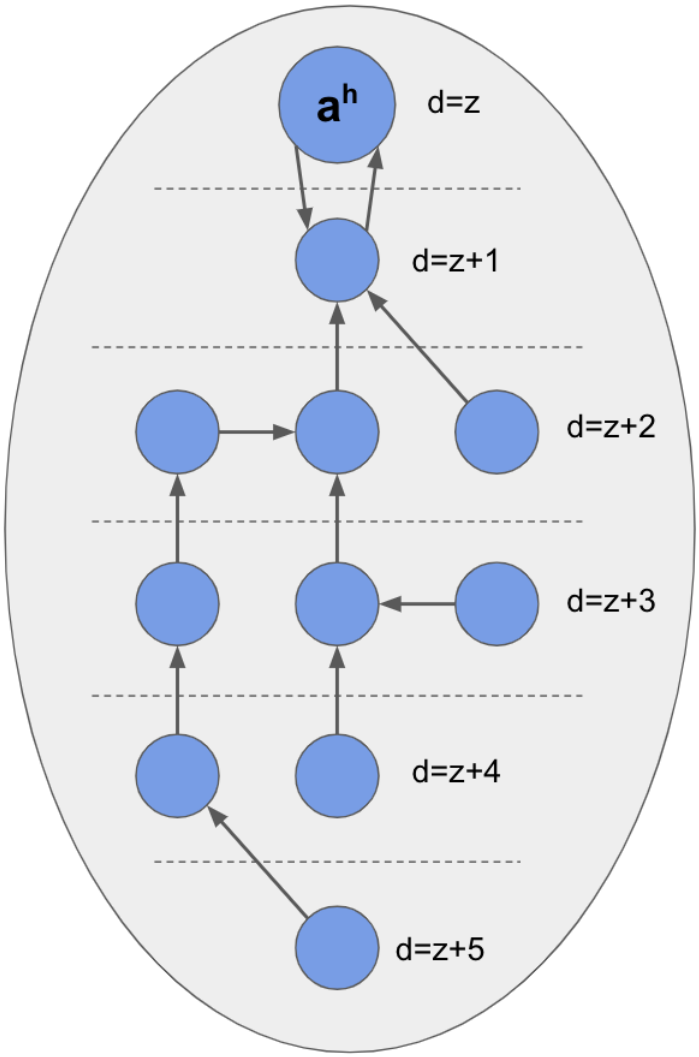

Figure B.2: Example subgraph Q_i formed by connecting classes through the minimum resistance edge exiting each class. There exists one class with minimal disagreement, a^h, and the edge leaving this class leads to a class with higher disagreement. All other edges in Q_i lead to classes with either the same or lower disagreement.

Proof: Consider a system that is influenced by some aggressive policy S that targets a. Within any minimum potential rooted tree in Q_i, all edges are between similar classes where the following transitions occur:

- $\{|Y|_y Y|_x Y|_y\} \rightarrow \{|Y|_y X|_x Y|_y\}$,
- $\{|X|_x Y|_y X|_x\} \rightarrow \{|Y|_y Y|_x Y|_y\}$,
- $\{|X|_y X|_x Y|_y\} \rightarrow \{|Y|_y Y|_x Y|_y\}$.

None of the transitions contain a segment that transitions from $|Y|_y$ to $|X|_y$. Thus, if a segment in a_i^h plays $|X|_y$, there are no classes in Q_i where that same segment plays $|Y|_y$ and therefore $\nexists\, i$ s.t. $a' \in Q_i : T_i = Y, a^h_{Q(i)} = X, a'_{Q(i)} = Y$. ■

Lemma B.1.9 *Consider a system that is influenced by some aggressive policy S that targets a. Then:*

$$\min_{u \in T_i, T', v \in T'} R(u \rightarrow v) = \min_{T', v \in T'} R(a_i^h \rightarrow v) \tag{B.19}$$

and

$$\min_{T', v \in T'} R(a_i^h \rightarrow v) = \min_{T' : d(T') < d(T_i), v \in T'} R(a_i^h \rightarrow v). \tag{B.20}$$

Proof: Consider a system that is influenced by some aggressive policy S that targets a. We will identify the minimum resistance edge leaving some rooted tree T_i in family Q_i and connecting to a class in another family. For all families $Q_i : a \notin Q_i$, $a_i^h \neq a$ and thus a_i^h contains at least one instance of $|X|y$. By transitioning to a similar class with the segment replaced with $|Y|y$, there is always edge leaving the family with resistance $1 + 2\alpha$. Because of this, any edges with resistance greater than $1 + 2\alpha$ do not need to be considered when searching for the minimum resistance edge leaving T_i. Any edges between classes with resistance $R\{|Y|y\} \rightarrow \{|X|_y\} > 1 + 2\alpha$ cannot be the minimum resistance edge leaving T_i.

The remaining edges that leave T_i that need to be considered are between similar classes where the following transitions occur:

- $\{|X|_x X|_y X|_x\} \rightarrow \{|X|_x Y|_y X|_x\}$,
- $\{|X|_y X|_x Y|_y\} \rightarrow \{|Y|_y Y|_x Y|_y\}$.

Both of these transitions contain an instance of $\{|X|_y\} \rightarrow \{|Y|_y\}$. By Lemma B.1.8, if the segment that becomes $\{|Y|y\}$ is not $\{|Y|y\}$ in a_i^h, then that transition at that segment does not occur in any edges in T_i. Additionally, a_i^h will contain the most instances of $\{|Y|y\}$ out of all classes in T_i. For any edge between similar classes that leaves T_i of the two aforementioned forms, there exists an edge with the same or less resistance leaving a_i^h where the same transition occurs. Therefore,

$$\min_{u \in T_i, T', v \in T'} R(u \rightarrow v) = \min_{T', v \in T'} R(a_i^h \rightarrow v). \tag{B.21}$$

Let the minimum resistance edge exiting T_i from a_i^h connect to a class in the minimum potential rooted tree T' in family Q'. Additionally, since the minimum resistance edge exiting T_i from a_i^h will be one of the two aforementioned types and both transitions lead to classes that either have lower disagreement than a_i^h or have an edge connecting them to a class with lower disagreement, $d(T') < d(T_i)$. Thus:

$$\min_{T', v \in T'} R(a_i^h \rightarrow v) = \min_{T': d(T') < d(T_i), v \in T'} R(a_i^h \rightarrow v). \tag{B.22}$$

■

Theorem B.1.1 *Under an aggressive policy based on a, the rooted tree with minimal stochastic potential is rooted in a.*

Proof: Consider a system attacked by an aggressive policy based on some action profile a. Let T be the rooted tree constructed by connecting all m of the minimum potential rooted trees T_i in families Q_i by the minimum resistance edges leaving each $a_i^h \neq a$ (see figure B.3). Since all edges that leave the T_i connect to trees with strictly lower disagreement, the tree ends up rooted in a, the class with minimal disagreement.

This tree has the minimum potential of all possible rooted trees because each sub-tree T_i is of minimal potential and they are connected using the minimum number of edges, $m-1$, where each edge is the minimal resistance edge leaving the tree T_i. Altering any edges to root the tree in any class that is not a will result in a tree with strictly greater potential. Because the minimum potential resistance tree is rooted in a, a is the sole strictly stochastic state of the system. ■

Lemma B.1.10 *Let a be an action profile with n segments and a' be an action profile with $n+2$ segments. Then, $R(a \to a') \geq 1-\alpha$.*

Proof: Let a be an action profile with n segments and a' be an action profile with $n+2$ segments. We will assume that the adversary sets are chosen to minimize the resistance along the path we outline. Any path from a to a' requires at least one transition of type $r(x,2,S_y) = 1+2\alpha$ or $r(y,2,S_x) = 1-\alpha$. If adversaries are assigned to sets in any other way, the resistance of these transitions can only stay the same or increase. Thus, $R(a \to a') \geq 1-\alpha$. ■

Lemma B.1.11 *Assume that an aggressive policy S can stabilize target action profile a within a system characterized by payoff gain α and adversarial budget γ. Let a' be an action profile that has the same number of segments as a. If a policy requires fewer adversaries than S, then there always exists some a' such that $R(a \to a') < 1-\alpha$.*

Proof: Assume that an aggressive policy S can stabilize target action profile a within a system characterized by payoff gain α and adversarial budget γ. Consider any policy $S' \neq S$ that uses fewer adversaries than S.

Let a' be an action profile that has the same number of segments as a. Because S' uses fewer adversaries than S, there will always exist a segment i in a' such that either $T_i = Y$ and $Q(i)$ is not targeted by a defensive y strategy or $T_i \in \{\hat{T}_1, \hat{T}_2, \ldots, \hat{T}_\lambda\}$ and $Q(i)$ is not targeted by a defensive x strategy. Then, it is always possible to construct a path from a to a'

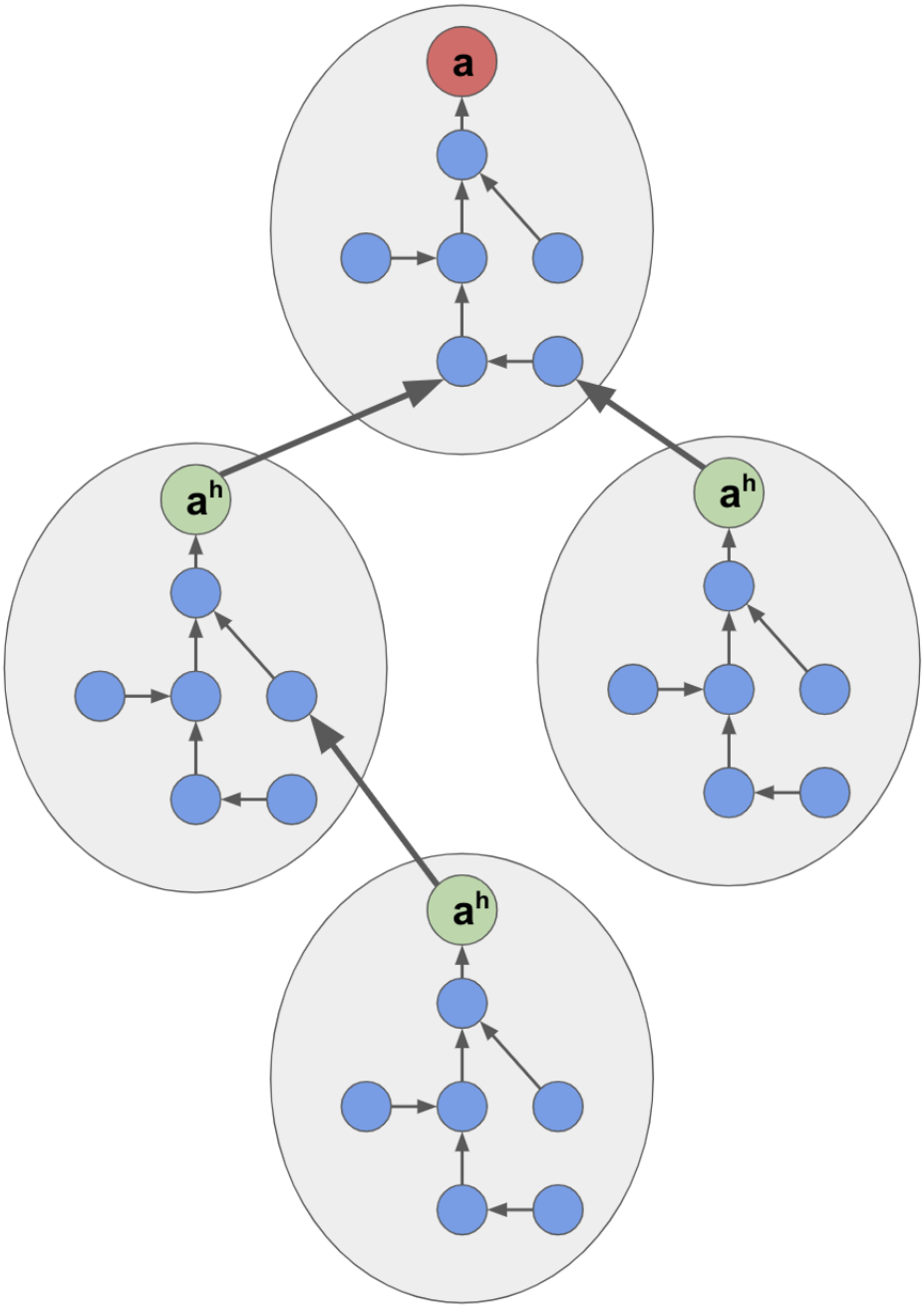

Figure B.3: Example minimum potential resistance tree rooted in a. Each family consists of classes connected by the minimum resistance edges leaving each node. The head of each family is connected to other families through the minimum resistance edge leaving its family. All edges either lead to a class with lower disagreement, or to a class that has an edge connected to a class with lower disagreement, resulting in a minimum potential tree rooted in a.

using only transitions of type $r(x,1,0) = \alpha$ and $r(y,1,0) = 0$. There will always be less than $\sum_{i=1}^{\lambda} \left(|\hat{T}_i| - 2\right)$ transitions of type $r(x,1,0)$, so by definition of λ, $R(a \to a') < 1 - \alpha$. ■

Theorem B.1.2 *If a policy uses fewer adversaries than an aggressive policy does to attempt to stabilize profile a, a will not be the sole stochastically stable state of the system.*

Proof: Assume that an aggressive policy S can stabilize target action profile a within a system characterized by payoff gain α and adversarial budget γ. Consider any policy $S' \neq S$ that uses fewer adversaries than S. Assume that we have a resistance tree rooted in a, and the adversary is using fewer adversaries than an aggressive policy calls for to attempt to stabilize a. By lemma B.1.11, there exists a path from a to some action profile a' with two less segments that has resistance less than $1-\alpha$. a' may or may not be a recurrent class. If a' is not a recurrent class, there exists some other action profile a'' that is a recurrent class with $R(a' \to a'') = 0$.

By lemma B.1.10, the path from a' to a contains an edge e with resistance of at least $1-\alpha$. It is possible to re-root the tree in a' (or a'' if a' is not a recurrent class) by eliminating edge e and routing a to a' (a'') along the minimum resistance path from a to a', where $R(a \to a') < 1-\alpha$. This new tree will no longer be rooted in a, and will have lower stochastic potential. Thus, a will not be the stochastically stable state of the system. ■

Corollary B.1.1 *If an action profile can be established as the sole stochastically stable state of a system, then an aggressive policy does so using the fewest number of adversaries.*

Lemma B.1.12 *Let k be the number of homogeneous Y segments and l be the number of homogeneous X segments of length 2 in some profile a. The minimum adversary budget γ needed to implement an aggressive policy based on a is then given by $\frac{2(k+l)}{n}$ when $\alpha < 0.5$ and $\frac{2k}{n}$ when $\alpha \geq 0.5$.*

Proof: Let k be the number of homogeneous Y segments and l be the number of x segments that are length 2 in some profile a. Under an aggressive policy, there needs to be two

adversaries per Y segment in a to ensure that a is a recurrent class. When $\alpha < 0.5$, there also needs to be two adversaries for each of the l X segments of length two. In both cases, there also needs to be two adversaries for each of the λ smallest segments playing X in a to ensure the resistances put forth in Lemma B.1.3 hold true. Because there will always be a Y segment for every X segment in a, $k \geq j$.

Consider a segment i. If i is targeted with a defensive y policy, adversaries will only be assigned to $Q(i)$ if the agents neighboring $Q(i)$ play x. If i is targeted with a defensive x strategy and $|Q(i)| > 2$, adversaries will only be assigned to $Q(i)$ if segment i is heterogeneous with one instance of two neighboring agents playing x. An aggressive policy requires $2k$ adversaries to ensure that defensive strategies can be deployed on all Y segments. There also needs to be 2λ adversaries to ensure that defensive strategies can be deployed on λ X segments. However, it is possible for strategies to "share" adversaries- if adversaries need to be deployed to an X segment of length greater than 2, there will necessarily be a neighboring Y segment that no longer needs an adversary deployed to the agent that borders the X segment to satisfy the requirements of a defensive y strategy (see Figure B.4). When $\alpha < 0.5$, there will always need to be two adversaries on each of the l X segments of length 2, then the remaining $\lambda - l$ X segments will be able to borrow from Y segments when necessary. In this case, the aggressive policy requires $2k + \lambda + l$ adversaries to implement defensive strategies. When $\alpha \geq 0.5$, X segments no longer need to be defended so the necessary number of adversaries needed to implement defensive strategies is given by $2k$.

In order to implement offensive strategies, an aggressive policy only needs one adversary. Thus, the total number of adversaries needed to implement a complete aggressive policy based on a is given by the number of adversaries required for defensive strategies plus one.

For any action profile a, another action profile a' with the same efficiency and m times the number of agents in a can be created by appending m action profiles a to each other. The minimum number of adversaries needed to implement an aggressive policy based on m is then

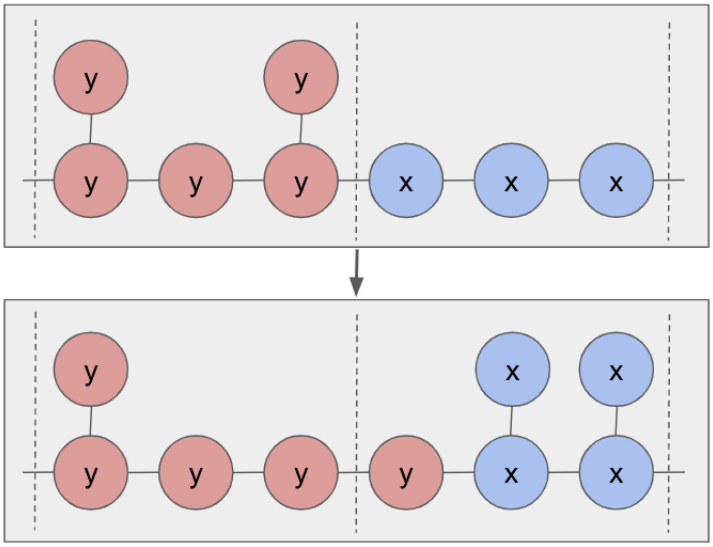

Figure B.4: Adversary allocation for two different action profiles. When adversaries are deployed to an X segment with length greater than two, a neighboring Y segment no longer needs an adversary on one of its edges.

given by $2m(k+l)+j+1$ when $\alpha < 0.5$ and $2mk+1$ when $\alpha \geq 0.5$. When $\alpha < 0.5$ the adversary budget needed to have the required number of adversaries is given by $\frac{2m(k+l)+j+1}{mn}$, and when $\alpha \geq 0.5$ it is given by $\frac{2mk+1}{mn}$. As m approaches infinity, this limit goes to $\frac{2(k+l)}{n}$. In the case where $\alpha \geq 0.5$, the limit instead goes to $\frac{2k}{n}$. ■

Lemma B.1.13 *It is impossible for any profile a that has a single agent playing x surrounded by agents playing y to be the stochastically stable state under the influence of a mobile intelligent adversary for any value of* $0 < \alpha < 1$.

Proof: Consider an action profile a that contains an instance of a single agent i playing x surrounded by agents playing y. Assume the system is influenced by the adversary set that maximizes the resistance of any unilateral deviation away from a involving i, $i \in S_x$. In this case, a unilateral deviation at i has a resistance of $[1+\alpha-2]_+ = 0$, meaning there exists a zero-resistance path leading out of a, thus disqualifying a from being a recurrent class and the sole stochastically stable state of the system. ■

Lemma B.1.14 *Consider a system attacked by an aggressive policy S based on some action profile a. If a has any segments i such that $T_i = Y$, $|Q(i)| < \left\lfloor \frac{2+\alpha}{1-\alpha} \right\rfloor + 1$, then a cannot be the sole stochastically stable state of the system.*

Proof: Consider an action profile a that contains a segment i such that $T_i = Y$ and $|Q(i)| < \left\lfloor \frac{2+\alpha}{1-\alpha} \right\rfloor + 1$. Let a' be a similar action profile where $a'_{Q(i)} = X$. There always exists a path between a and a' consisting of $|Q(i)| - 1$ transitions of type $r(y, 1, y)$. Then, $R(a \to a') < (1-\alpha)\left\lfloor \frac{2+\alpha}{1-\alpha} \right\rfloor < 1 + 2\alpha$. By Lemma $B.1.4$, $R(a' \to a) = 1 + 2\alpha$. Let T be the minimum resistance tree rooted in a. There always exists a tree T' rooted in a' obtained by removing the edge (a', a) and replacing it with (a, a') that has lower stochastic potential than T. Thus, a cannot be the sole stochastically stable state of the system. ■

We are now ready to pose the optimization to find the minimum efficiency profile a that can be stabilized in a system characterized by payoff gain α and adversary budget γ. By Lemmas B.1.13 and B.1.14, all X segments must be length 2 or greater and all Y segments must be length $\left\lfloor \frac{2+\alpha}{1-\alpha} \right\rfloor + 1$ or more. By B.1.12, if $\alpha < 0.5$ there needs to be two adversaries for each Y segment of any length and X segment of length two in a. If $\alpha \geq 0.5$, there only needs to be two adversaries for every Y segment in a. Then, the optimization can be posed as a modified version of (4.1) with adjusted surpluses s_k. ■

B.2 Proof of Theorem 4.1.6

This proof follows similarly to the proof of Theorem 4.1.3. Upon relaxing the integer constraint in Theorem 4.1.6, it is possible to solve for a closed form expression for the minimal efficiency. The major difference is that in the mobile intelligent case, the necessary number of adversaries needed to stabilize a segment is a constant instead of a linear function of the

lengths of the X and Y segments. Solving for efficiency we obtain:

$$\eta(\gamma, \alpha, \ell_x) = \frac{c - 2\gamma + \alpha\gamma(\ell_x - 1)}{c(1+\alpha)}, \tag{B.23}$$

where $c = 2$ if $\ell_x > \frac{1}{\alpha}$ and $c = 4$ otherwise. For both of these cases, (B.23) is minimized when ℓ_x is minimized. For the first regime, the minimum value of ℓ_x is $\frac{1}{\alpha}$ while in the second it is 2. By then solving (B.23) for the minimum efficiency in both cases, the case where $\ell_x \geq \frac{1}{\alpha}$ will always have the lower efficiency, given by:

$$\eta(\gamma, \alpha) = \frac{2 - \gamma(1-\alpha)}{2(1+\alpha)}. \tag{B.24}$$

When $\alpha > \frac{1}{2}$, ℓ_x cannot be less than 2. In this regime, the minimum efficiency is given by:

$$\eta(\gamma, \alpha) = \frac{2 - 2\gamma + \alpha\gamma}{2(1+\alpha)}. \tag{B.25}$$

Thus, the minimum efficiency attainable as a function of η and α is given by:

$$\eta(\gamma, \alpha) = \begin{cases} \frac{2-\gamma(1-\alpha)}{2(1+\alpha)} & \text{if } \alpha < \frac{1}{2}, \\ \frac{2-2\gamma+\alpha\gamma}{2(1+\alpha)} & \text{else.} \end{cases} \tag{B.26}$$

In addition to the minimum efficiency equations, there are also "saturation" limits for γ where increasing γ past some critical point does not decrease efficiency due to length limitations on X and Y segments. Define $c(\alpha)$ as the saturation limit when minimum efficiency is given by (B.24) and $b(\alpha)$ when it is given by (B.25). In the regime $0 < \alpha \leq \frac{1}{4}$, the minimum lengths of X and Y segments are respectively $\frac{1}{\alpha}$ and 3. For $\frac{1}{4} < \alpha \leq \frac{1}{2}$, the minimum lengths of X and Y segments are respectively $\frac{1}{\alpha}$ and $\frac{2+\alpha}{1-\alpha}$. Finally, for $\frac{1}{2} < \alpha \leq 1$, the minimum lengths of X and Y segments are respectively 2 and $\frac{2+\alpha}{1-\alpha}$. When these limitations are applied to (B.24) and

(B.25), it is possible to calculate $b(\alpha)$ and $c(\alpha)$ as follows:

$$b(\alpha) = \begin{cases} \frac{4}{5} & \alpha \leq \frac{1}{4}, \\ \frac{4(1-\alpha)}{4-\alpha}, & \end{cases}$$

$$c(\alpha) = \begin{cases} \frac{2\alpha}{1+3\alpha} & \alpha \leq \frac{1}{4}, \\ \frac{2\alpha(1-\alpha)}{1+\alpha+\alpha^2} & \frac{1}{4} < \alpha \leq \frac{1}{2}, \\ \frac{2(1-\alpha)}{4-\alpha} & \text{else.} \end{cases}$$

By applying these saturation values to (B.24) and (B.25) and taking the minimum efficiency of the two functions in each of the three regimes, we can determine the minimum efficiency as a function of α and γ. In the region where $b(\alpha) < \gamma < (\alpha)$, the minimum efficiency is achieved by using a combination of segments using the saturation lengths used to determine $b(\alpha)$ and $c(\alpha)$ to get a line that connects the minimum efficiency of the $b(\alpha)$ regime to the minimum efficiency of the $c(\alpha)$ regime. ■

Appendix C

Oblivious Adversaries

C.1 Proof of Theorem 4.2.1

If $k_x > 0$, there exist influence sets which induce equilibria that have strictly higher efficiency than $\vec{y}$. For $k_x \geq k_y$, these sets are created by alternating the indexes assigned to each influence set such that $\{1, 3, \ldots, 2k_y - 1\} \in S_x$ and $\{2, 4, \ldots, 2k_y\} \in S_y$, then assigning the remaining adversaries to S_x. The argument for $k_y > k_x$ is similar. Because a heterogeneous action profile can never be stochastically stable if it strictly alternates between x and y (by Lemma A.1.2), this means that the adversary can never guarantee that a heterogeneous action profile is induced, leaving only $\vec{y}$ and implying that $k_y = k$. ■

C.2 Proof of Theorem 4.2.2

The proof of Theorem 4.2.2 proceeds in two steps. First, Lemma C.2.1 shows that a random adversary cannot gain anything by influencing any agent to play x. Next, we combine this fact with results from [18] to complete the proof.

Lemma C.2.1 *When $\alpha < 1/2$, the solution to* (3.8) *has $k_x = 0$ and $k_y = k$.*

Proof: Regardless of what the adversary chooses, the only possible stochastically-stable action profiles are $\vec{x}$ and $\vec{y}$; this essentially follows from the fact that they are the only Nash equilibria of the nominal game. Because of this, the adversary's only viable option is to maximally target the $\vec{y}$ profile, which is accomplished by choosing $k_x = 0$ and $k_y = k$. ■

When the adversary chooses $k_x = 0$, the problem reduces to that of Theorem 5 in [18], in which an adversary randomly influences subsets of N with action y. In that setting, it is shown that for any $k \in \{1, \ldots, n-1\}$, joint action $\vec{y}$ is strictly stochastically stable if and only if $\alpha < 1/2$, and that if $\alpha \geq 1/2$, joint action $\vec{x}$ is stochastically stable (strict if $\alpha > 1/2$). The efficiency of $\vec{y}$ is $1/(1+\alpha)$, yielding (4.7). ■

Bibliography

[1] S. Martínez, J. Cortés, and F. Bullo, *Motion Coordination with Distributed Information*, *Control Systems Magazine* **27** (2007), no. 4 75–88.

[2] R. Olfati-Saber, J. A. Fax, and R. M. Murray, *Consensus and cooperation in networked multi-agent systems*, *Proceedings of the IEEE* **95** (2007), no. 1 215–233, [arXiv:1009.6050].

[3] M. Zhu and S. Martínez, *Distributed coverage games for mobile visual sensors (I): Reaching the set of Nash equilibria*, in *Proceedings of the IEEE Conference on Decision and Control*, pp. 169–174, 2009. arXiv:1002.0367.

[4] V. Ramaswamy and J. R. Marden, *A sensor coverage game with improved efficiency guarantees*, in *Proceedings of the American Control Conference*, vol. 2016-July, pp. 6399–6404, 2016.

[5] A. Jadbabaie, J. Lin, and S. A. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules.*, *Transactions on Automatic Control* **48** (2003), no. 6 988–1001.

[6] N. Li and J. R. Marden, *Designing games for distributed optimization*, *IEEE Journal on Selected Topics in Signal Processing* **7** (2013), no. 2 230–242.

[7] J. R. Marden, H. P. Young, and L. Y. Pao, *Achieving pareto optimality through distributed learning*, *SIAM Journal on Control and Optimization* **52** (2014), no. 5 2753–2770.

[8] J. R. Marden and J. S. Shamma, *Game Theory and Distributed Control*, in *Handbook of Game Theory Vol. 4* (H. Young and S. Zamir, eds.). Elsevier Science, 2014.

[9] L. Pavel, *Game Theory for Control of Optical Networks*. Springer Science & Business Media, 2012.

[10] D. H. Wolpert and K. Tumer, *Optimal Payoff Functions for Members of Collectives*, *Advances in Complex Systems* **04** (2001), no. 02n03 265–279.

[11] R. D. McKelvey and T. R. Palfrey, *Quantal response equilibria for normal form games*, *Games and Economic Behavior* **10** (1995), no. 1 6–38.

[12] L. E. Blume, *The Statistical Mechanics of Best-Response Strategy Revision*, *Games and Economic Behavior* **11** (1995), no. 2 111–145.

[13] C. Alós-Ferrer and N. Netzer, *The logit-response dynamics*, *Games and Economic Behavior* **68** (2010), no. 2 413–427.

[14] M. Kearns, M. L. Littman, and S. Singh, *Graphical Models for Game Theory*, *Proceedings of the 17th conference on Uncertainty in artificial intelligence* (2001), no. Owen 253–260, [arXiv:1301.2281].

[15] A. Montanari and A. Saberi, *The spread of innovations in social networks.*, *Proceedings of the National Academy of Sciences of the United States of America* **107** (2010), no. 47 20196–20201.

[16] D. Shah and J. Shin, *Dynamics in Congestion Games*, *Proc. ACM SIGMETRICS'10* **38** (2010) 107.

[17] H. P. Borowski and J. R. Marden, *Understanding the Influence of Adversaries in Distributed Systems*, in *IEEE Conference on Decision and Control (CDC)*, pp. 2301–2306, 2015.

[18] P. N. Brown, H. P. Borowski, and J. R. Marden, *Security Against Impersonation Attacks in Distributed Systems*, *IEEE Transactions on Control of Network Systems* (2018) [arxiv.org/abs/1710.08500].

[19] Y. Lim and J. Shamma, *Robustness of stochastic stability in game theoretic learning*, in *American Control Conference (ACC)*, pp. 6160–6165, 2013.

[20] J. R. Marden, G. Arslan, and J. S. Shamma, *Joint strategy fictitious play with inertia for potential games*, *IEEE Transactions on Automatic Control* **54** (2009), no. 2 208–220.

[21] B. Pradelski and H. P. Young, *Learning Efficient Nash Equilibria in Distributed Systems*, *Games and Economic Behavior* **75** (2012), no. 2 882–897.

[22] H. P. Young, *The Evolution of Conventions*, *Econometrica* **61** (1993), no. 1 57–84.

[23] D. Monderer and L. S. Shapley, *Potential Games*, *Games and Economic Behavior* **14** (1996), no. 1 124–143.